An exploration of the protein component of scent marks from a range of mammalian species using proteomics approaches

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

Grace May Loxley

September 2019

Acknowledgements

I would like to start by thanking my supervisors, Professor Rob Beynon and Professor Jane Hurst, for their incredible support, guidance and enthusiasm throughout both my master's degree and my PhD for which I am extremely grateful. I would also like to thank all fellow members of the Centre for Proteome Research for their kindness and support, particularly Dr Guadalupe Gómez-Baena and Dr Philip Brownridge, for your training and mentorship.

I would also like to thank Holly Coombes, Alexandra Jebb, and all other members of the Mammalian Behaviour and Evolution Group at the University of Liverpool in their assistance in obtaining samples as well as their expertise in mammalian behaviour. I would additionally like to thank my collaborators at Victoria University of Wellington for the opportunity to work on such an exciting project.

Becka, Gemma and Sam; I am so grateful for the cheese and wine (mostly wine) nights, endless chatter and ridiculous conversations that we have had, you are truly friends for life. I would also like to thank Amber and Sophie for the countless evenings of food, gossip and laughter- I hope our tour of Liverpool restaurants is far from over. To Hannah, I would like to thank you for your no-nonsense advice, laughter and lifetime friendship- I'm sorry I moved so far south again. To Max, I would like to thank you for your friendship and encouragement, and for tolerating so many months of me commandeering your living room and WiFi.

Thank you to my parents, for your endless patience and encouragement. From finding a flat for me while I was writing, to spending hours of your life driving up and down the M6 over the last 8 years, you have always gone the extra mile (or 300) to support me.

And lastly to Aidan, for your combination of unfailing support, enthusiasm, rationality, variable culinary skills, and patience. For tolerating my delusional assertions that in 'just a few more weeks' I would have finished writing and would be free to enjoy life with you again, and for your unvarying positivity and encouragement, without which this process would have been unimaginably harder; thank you.

Abstract

An exploration of the protein components of scent marks from a range of mammalian species using proteomics approaches

Whilst semiochemistry is traditionally associated with volatile cues, the discovery of the major urinary proteins (MUPs) of the house mouse, and their demonstrable ability to convey information such as identity, kinship and mating availability, have identified proteins as an important tool for communication. Whilst protein-mediated scent signalling is understood relatively well in the house mouse and the Norway rat, little is known regarding the roles of proteins in chemical signalling beyond these species. The protein content of scent marks from three species, the closely related bank vole (*Myodes glareolus*) and field vole (*Microtus agrestis*), and the much more distantly related marsupial, the brushtail possum (*Trichosurus vulpecula*), were investigated.

Overall protein content was initially assessed using polyacrylamide gel electrophoresis and electrospray ionisation-mass spectrometry of intact proteins. Prominent proteins of interest were separated by anion exchange chromatography and sequenced *de novo* using liquid chromatography allied to a tandem mass spectrometer. Total proteome analyses were performed by tryptic proteolytic cleavage followed by liquid chromatography-tandem mass spectrometry. Proteins were identified by cross-species matching and quantified using a label-free approach.

Previous investigation into urinary protein expression in the bank vole, *Myodes glareolus*, identified three odorant-binding proteins (OBPs). The discovery of another OBP, glareosin, is described. Glareosin was determined as the single most abundant protein present in male bank vole urine during the breeding season, and the elucidation of its sequence and structure was published in 2017. Additional work continued to explore the total protein content, identifying additional OBP-like proteins at the peptide level in urine and scent marks of both sexes, including previously identified OBPs, although no corresponding intact masses were found.

Behavioural research into the field vole (*Microtus agrestis*) has concentrated on its unusual dynamic population cycles, but olfactory communication remain largely unevaluated. Three abundant male-specific sequence variants, homologous to bank vole glareosin, were identified from male breeding season urine and sequenced *de novo*. Other proteins, including another OBP-like protein, a putatively glycosylated MUP-like protein, and a lipocalin-11-like protein were partially sequenced from urine and scent marks of both

species. Global proteome analysis indicated further unidentified heterogeneity of OBP proteins at the peptide level, indicating a far more complex protein landscape within field vole scent marks when compared to the bank vole.

In New Zealand, the status of the brushtail possum, *Trichosurus vulpecula*, as an invasive pest species has ignited a concerted effort to eradicate the species, despite a deficit in research into marsupial behaviour. Following proteome analyses, a glycosylated lipocalin was identified in the urine of both sexes, and sequenced *de novo*. Phylogenetic analysis with a range of mammalian lipocalins identified the novel sequence within a new clade of lipocalins unique to marsupials, suggesting that the marsupial lineage diverged prior to the establishment of described lipocalin classes in the placental mammals.

The following investigations are three examples of the diversity of scent mark protein expression between both closely- and distantly-related species, and highlight the lack of understanding in this area. However, due to the evolutionary pressure placed on proteins with capacity to influence mate choice and sexual selection, a proteome approach reliant on identification by database matching is difficult, particularly for those species without a comprehensive genome, which will remain a rate-limiting factor until a wider range of genomes are available.

Contents

Acknowledgements i
Abstract ii
List of Figures ix
List of Tables xv
Supplementary Material xvi
1 Introduction 1
1.1 Mammalian chemical scent communication 2
1.1.1 Social organisation 2
1.1.2 Territorial marking 4
1.1.3 Juvenile care 5
1.1.4 Sexual selection 6
1.2 The mammalian anatomy of pheromone detection 9
1.2.1 Accessory olfactory system 10
1.3 Chemical diversity of pheromones 12
1.3.1 Volatile semiochemicals 12
1.3.2 Involatile semiochemicals 15
1.4 Mammalian Proteins in Scent-Mediated Communication 16
1.4.1 Salivary androgen-binding protein (ABP) 17
1.4.2 Exocrine gland-secreting peptides (ESPs) 17
1.4.3 Major Histocompatibility Complex (MHC) Peptides 18
1.4.4 Lipocalins 19
1.5 Methodology of scent mark analysis 32
1.5.1 Assessment of sample complexity 33
1.5.2 Protein identification 34
1.5.3 Sequencing de novo 37
1.5.4 Protein quantification 42
1.5.5 Glycoprotein analysis 44
1.6 Scope of this thesis 46
Aims & Objectives 48
2 Experimental Strategy 50
2.1 Abstract 50
2.2 Overview 51
2.3 Sample Collection 53
2.3.1 Bank vole and field vole housing conditions 53
2.3.2 Bank vole and field vole urine collection 53
2.3.3 Bank vole and field vole scent mark collection 53
2.3.4 Bank vole and field vole preputial gland collection 54
2.3.5 Bank vole and field vole bladder urine collection 54
2.4 Initial Assessment 54
2.4.1 Protein assay 54
2.4.2 Creatinine assay 54
2.4.3 Polyacrylamide gel electrophoresis 54
2.5 Protein digestion 55
2.5.1 In-gel digestion 55
2.5.2 In-solution digestion 55
2.6 Mass Spectrometry Techniques 56
2.6.1 Electrospray-mass spectrometry of intact proteins 56
2.6.2 MALDI-ToF Mass Spectrometry 56
2.6.3 Tandem mass spectrometry 57
2.7 Data analysis 59

2.7.1 Database searching 59
2.7.2 De novo sequencing 59
2.7.3 Label-free quantification 60
2.8 Metabolic labelling 61
2.9 Anion exchange chromatography 61
2.10 Clean-up and concentration of proteins and peptides 62
2.10.1 Strataclean capture 62
2.10.2 Vivaspin® membrane ultrafiltration concentration 62
2.10.3 Desalting columns 62
2.11 Sequence comparison 62
2.11.1 BLAST searching 62
2.11.2 Multiple Sequence Alignment 63
2.11.3 Phylogenetic Analysis 63
2.11.4 Structural Homology Modelling 63
2.11.5 Statistical Analysis 64
3 Characterisation of the protein content of scent secretions in the bank vole, *Myodes glareolus* 65
Abstract 65
Contributions 66
Published Paper: 67
3.1 Introduction 68
3.2 Aims & Objectives 70
3.3 Methods 71
3.3.1 Sample collection 71
3.3.2 Protein output 71
3.3.3 Polyacrylamide gel electrophoresis 71
3.3.4 In-gel digestion 71
3.3.5 In-solution proteolysis 71
3.3.6 Edman degradation 71
3.3.7 MALDI-ToF mass spectrometry 72
3.3.8 Electrospray ionization mass spectrometry of intact proteins 72
3.3.9 Tandem mass spectrometry 72
3.3.10 Database searching 73
3.3.11 Label-free quantification 73
3.3.12 Metabolic labelling 73
3.3.13 Anion Exchange Chromatography 74
3.3.14 Protein sequence analysis: 74
3.3.15 Phylogenetic analysis: 74
3.3.16 Homology modelling: 75
3.4 Results 76
3.4.1 Preliminary work 76
3.4.2 Sequencing de novo 83
3.4.3 Discrimination of leucine and isoleucine residues 86
3.4.4 Phylogenetic analysis 96
3.4.5 Structural homology modelling 97
3.4.6 Further identification of proteins in bank vole urine 100
3.4.7 Global proteomics of bank vole urine 107
3.4.8 Preliminary investigation into the protein content of bank vole scent marks 118
3.4.9 In-gel proteome analysis of bank vole scent marks 123
3.4.10 Global proteome analysis of bank vole scent marks and bladder urine 130
3.5 Discussion 142

3.6 Supplementary 145
3.6.1 Published paper 145
3.6.2 Individual intact protein mass spectra 145
3.6.3 Glareosin sequencing fragment ion data (from paper supplementary) 145
3.6.4 Glareosin heavy leucine data 145
3.6.5 Multiple sequence alignment for phylogenetic analysis (from paper supplementary) 145
3.6.6 Bladder urine intact protein mass spectra 145
3.6.7 Female scent mark intact protein mass spectra 145
4 Characterisation of the urinary protein content in the field vole, *Microtus agrestis* .. 146
Abstract 146
Contributions 146
4.1 Introduction 147
4.2 Aims & Objectives 149
4.3 Methods 150
4.3.1 Housing conditions 150
4.3.2 Urine collection 150
4.3.3 Scent mark collection 150
4.3.4 Protein assay 150
4.3.5 Creatinine assay 150
4.3.6 Polyacrylamide gel electrophoresis 150
4.3.7 In-gel digestion 150
4.3.8 In-solution digestion 150
4.3.9 Electrospray-mass spectrometry of intact proteins 150
4.3.10 Tandem mass spectrometry 150
4.3.11 Database searching 150
4.3.12 De novo sequencing 150
4.3.13 Label-free quantification 150
4.3.14 Anion Exchange Chromatography 151
4.3.15 Desalting columns 151
4.3.16 BLAST searching 151
4.3.17 Multiple Sequence Alignment 151
4.3.18 Structural Homology Modelling 151
4.3.19 Statistical Analysis 151
4.4 Results 152
4.4.1 Initial Assessment 152
4.4.2 Identification of major male field vole urinary proteins 158
4.4.3 Primary sequence-level heterogeneity of field vole glareosin 168
4.4.4 Identification of major urinary proteins in the female field vole 180
4.4.5 Identification of field vole urinary proteins separated by PAGE and digested in-gel. 183
4.4.6 Global proteomics of mature field vole urine 185
4.4.7 Urine from sexually immature field voles 187
4.4.8 A preliminary exploration of scent mark protein content in the field vole, Microtus agrestis 195
4.5 Discussion 212
4.6 Supplementary 216
4.6.1 Intact mass profiles of urine from mature field voles. 216
4.6.2 Fragment ion spectra of peptides used to sequence an initial draft of field vole glareosin. 216
4.6.3 Sequence coverage of odorant binding proteins from analysis of pooled female field vole urine. 216

4.6.4 Alignment of major urinary proteins to support preliminary sequencing of a field vole major urinary protein. 216
4.6.5 Intact mass profiles of juvenile field vole urine. 216
5 Characterisation of the urinary protein content in the New Zealand brushtail possum, *Trichosurus vulpecula* 217
Abstract 217
Contributions 217
5.1 Introduction 218
5.2 Aims & Objectives 221
5.3 Methods 222
5.3.1 Urine sampling 222
5.3.2 Protein and creatinine assays 222
5.3.3 SDS-PAGE and Native PAGE analysis 222
5.3.4 Anion exchange chromatography 222
5.3.5 Intact protein mass measurement 223
5.3.6 In-gel digestion and MALDI-ToF analysis 223
5.3.7 Protein desalting and concentration 223
5.3.8 In-solution digestion and LC-MS/MS 224
5.3.9 Peptide data searches 224
5.3.10 Label-free quantification 225
5.3.11 Protein deglycosylation 225
5.3.12 Discrimination of leucine and isoleucine by metabolic labelling: 225
5.3.13 Phylogenetic analysis 226
5.3.14 Homology modelling: 226
5.4 Results 228
5.4.1 Preliminary investigation into urinary protein content 228
5.4.2 Protein identification 241
5.4.3 Protein Purification 247
5.4.4 Protein sequencing 249
5.4.5 Deglycosylation 252
5.4.6 Purification, deglycosylation and sequencing of a urinary protein in female brushtail possum urine. 267
5.4.7 Distinction of leucine & isoleucine residues using isotopic labelling 270
5.4.8 Sequence analysis of the novel brushtail possum protein 274
5.4.9 Structural Homology Modelling of vulpeculin 277
5.4.10 Proteome analysis 280
5.5 Discussion 289
5.6 Characterisation of the urinary protein content in the New Zealand brushtail possum, *Trichosurus vulpecula:* Supplementary information 294
5.6.1 Intact mass profiles of individual samples 294
5.6.2 MALDI-ToF peptide mass fingerprints from tryptic in-gel digestion 294
5.6.3 Homologous sequences to the trichosurin-like protein from M. domestica 294
5.6.4 Fragment ion spectra for sequencing 294
5.6.5 Fragment ion spectra for determination of leucine and isoleucine residues 294
5.6.6 Table of accession numbers for multiple sequence alignment 294
5.6.7 Multiple sequence alignment of lipocalins 294
6 General Discussion 295
6.1 Summary of results and implications in behaviour and social structure 295
6.1.1 Myodes glareolus 296
6.1.2 Microtus agrestis 298

6.1.3 Trichosurus vulpecula 301
6.2 Methodology 305
6.3 Summary 308
7 References 310

List of Figures

Figure 1.1 | *Structure of the nasal cavity across the animal kingdom.* 9
Figure 1.2 | *Structure of lipocalins.* 21
Figure 1.3 | *Classes of lipocalins.* 22
Figure 1.4 | *Sequence conservation in the lipocalin family.* 23
Figure 1.5 | *Structure of lipocalins.* 24
Figure 1.6 | *Structure of lipocalins.* 25
Figure 1.7 | *Phylogeny of lipocalin sequences in rodents.* 26
Figure 1.8 | *Structure of b- and y-ions.* 38
Figure 1.9 | *Example fragment ion spectra.* 39
Figure 1.10 | *Sequencing* de novo *using overlapping peptides.* 41
Figure 1.11 | *Action of PNGase F.* 45
Figure 2.1 | *Experimental strategy*. 52
Figure 3.1 | *Analysis of bank vole urinary protein output.* 77
Figure 3.2 | *Exploring consistency in bank vole urinary protein output.* 80
Figure 3.3 | *Exploring continuity in wild bank vole urinary protein output.* 82
Figure 3.4 | Complete amino sequence of the novel bank vole urinary protein. 84
Figure 3.5 | Metabolic labelling strategy to discern leucine and isoleucine residues..... 87
Figure 3.6 | Determining the incorporation rate of heavy leucine-labelled peptides. 89
Figure 3.7 | *Dietary incorporation of 5,5,5-d3 leucine over time*. 90
Figure 3.8 | Determining leucine and isoleucine residues from fragment ion spectra..... 91
Figure 3.9 | Proportion of differentially labelled peptides during 5,5,5-d3 leucine incorporation with two ambiguous leu/ile sites..... 93
Figure 3.10 | *Resolution of leucine/isoleucine residues in a missed cleavage peptide.*..... 94
Figure 3.11 | Resolution of leucine and isoleucine by metabolic labelling. 95
Figure 3.12 | *Phylogenetic tree of glareosin-related sequences.* 96
Figure 3.13 | *Predicted three-dimensional structure of glareosin.* 99
Figure 3.14 | In-gel identification of major protein bands separated by SDS and native PAGE..... 103
Figure 3.15 | In-gel identification of major protein bands separated by SDS and native PAGE..... 105
Figure 3.16 | In-gel identification of major protein bands separated by native PAGE. 106
Figure 3.17 | Summary of protein identifications using cross-species matching against a multiple-species database and a single-species database..... 109

Figure 3.18 | *Sequence coverage of odorant binding proteins.* 114
Figure 3.19 | Result-based filtering of feature vectors for label-free quantification in PEAKS™. 116
Figure 3.20 | *Label-free quantification.* 117
Figure 3.21 | *Label-free quantification.* 117
Figure 3.22 | SDS-PAGE and Bradford assay of the protein component of bank vole scent marks 119
Figure 3.23 | *Intact mass analysis of bladder urine*. 120
Figure 3.24 | Intact mass analysis of bank vole scent marks. 122
Figure 3.25 | Identification of proteins in female bank vole scent marks by in-gel digestion and LC-MS/MS analysis. 125
Figure 3.26 | Identification of proteins in male bank vole scent marks by in-gel digestion and LC-MS/MS analysis. 126
Figure 3.27 | Identification of proteins in male bank vole preputial gland secretion by in-gel digestion and LC-MS/MS analysis. 127
Figure 3.28 | Identification of protein in male bank vole bladder urine, recovered urine and scent marks by in-gel digestion and LC-MS/MS. 128
Figure 3.29 | Sequence coverage of glareosin from peptide data of male bladder urine samples. 132
Figure 3.30 | OBP sequence coverage from in-solution digestion and LC-MS/MS analysis of female scent mark samples. 134
Figure 3.31 | OBP sequence coverage from de novo tag spectra generated from LC-MS/MS analysis of peptides retrieved from in-solution digestion of female scent marks. 136
Figure 3.32 | Sequence coverage of glareosin, OBPs and aphrodisin-like protein (C.griseus) from peptide data generated from in-solution digestion of male scent mark samples (n = 5). 139
Figure 4.1 | Analysis of the protein output of field vole urine samples. 154
Figure 4.2 | *Intact mass analysis of male field vole urinary proteins*. 155
Figure 4.3 | *Intact mass analysis of female field vole urinary proteins* 156
Figure 4.4 | Intact mass profiling of pooled male and female field vole urine samples. 157
Figure 4.5 | Peptide coverage of glareosin and the odorant binding proteins (OBPs) when searching peptides generated from proteolytic digestion of proteins in pooled male field vole urine. 160

Figure 4.6 | Peptide coverage of major urinary protein-like sequences from the golden hamster when searching peptides generated from proteolytic digestion of proteins in pooled male field vole urine. 161
Figure 4.7 | Preliminary field vole glareosin sequence built from overlapping peptides generated in PEAKS™. 162
Figure 4.8 | *Peptide coverage of initial field vole glareosin sequence.* 163
Figure 4.9 | *Sites of possible heterogeneity in field vole glareosin.* 166
Figure 4.10 | Separation of proteins from pooled male field vole urine by anion exchange chromatography. 169
Figure 4.11 | *Intact mass analysis of anion exchange separated proteins from pooled male field vole urine.* 170
Figure 4.12 | Protein profiles of male field vole urine samples used to confirm presence of point mutations. 173
Figure 4.13 | Peptide map of field vole glareosin variants in a male sample with a protein profile dominated by a single mass peak, 17168 Da. 176
Figure 4.14 | Peptide maps of field vole glareosin variants in a male urine sample with four main peaks; 17138, 17168, 17236 and 17252 Da. 178
Figure 4.15 | Location and structure of proposed amino acid substitutions causing protein-level heterogeneity in field vole glareosin. 180
Figure 4.16 | Identification and sequence coverage of a major urinary protein in pooled female field vole urine. 182
Figure 4.17 | In-gel identification of protein bands from SDS (top) and native (bottom) PAGE analysis of urine pooled from mature male and female field voles. 184
Figure 4.18 | Label-free quantification of urinary proteins in mature field voles. 186
Figure 4.19 | Initial assessment of protein output in the urine of juvenile field voles. 188
Figure 4.20 | Intact mass analysis of urinary proteins from juvenile field voles. 189
Figure 4.21 | Label-free quantification of proteins detected in juvenile field vole samples. 191
Figure 4.22 | Sequencing a draft of field vole OBP3 from multiple protease digestion of two juvenile female field vole urine samples. 193
Figure 4.23 | Initial sequencing of field vole OBP3 from analysis of juvenile field vole urine. 194
Figure 4.24 | Initial assessment of the protein content of field vole scent marks. 195
Figure 4.25 | *Intact mass analysis of field vole scent marks.* 197

Figure 4.26 | Identification of major female field vole scent mark proteins separated by molecular weight using SDS-PAGE. 199
Figure 4.27 | Identification of major male field vole scent mark proteins separated by molecular weight using SDS-PAGE. 200
Figure 4.28 | Peptide coverage of lipocalin 11-like proteins observed in male scent marks. 203
Figure 4.29 | Sequence coverage of lipocalin sequences identified in female scent marks. 204
Figure 4.30 | Sequence coverage of lipocalin sequences identified in male scent marks. . 206
Figure 4.31 | Alignment of rodent secretoglobin sequences with manually selected peptides sequenced in PEAKS™ from analysis of field vole scent marks. 209
Figure 4.32 | Peptide evidence of the secretoglobin family in female field vole scent marks. 210
Figure 4.33 | Peptide evidence of the secretoglobin family in male field vole scent marks. 211
Figure 5.1 | *Investigation into the urinary protein content of T. vulpecula by SDS-PAGE and overall protein content.* 228
Figure 5.2 | *Investigation into the urinary protein content of T. vulpecula by intact protein analysis.* 230
Figure 5.3 | *Initial investigation into the urinary protein content of T. vulpecula by intact protein analysis.* 231
Figure 5.4 | *Investigation into the urinary protein content of T. vulpecula by intact protein analysis: examples of variation in the deconvoluted mass spectrum.* 233
Figure 5.5 | *Investigation into the urinary protein content of T. vulpecula by intact protein analysis: examples of m/z spectra varying in spectral quality.* 234
Figure 5.6 | *Investigation into the protein content of replicate urine samples from the same individual.* 235
Figure 5.7 | *Intact protein analysis of ten urine samples from five male brushtail possums.* 236
Figure 5.8 | *Intact protein analysis of ten urine samples from five male brushtail possums.* 237
Figure 5.9 | *MALDI-ToF analysis of in-gel digested protein separated by SDS-PAGE and resolving at approximately 23-25 kDa from male brushtail possum urine.* 239
Figure 5.10 | *MALDI-ToF analysis of in-gel digested protein separated by SDS-PAGE and resolving at approximately 23-25 kDa from female brushtail possum urine.* 240

Figure 5.11 | *Protein identification.* 242
Figure 5.12 | *Protein Identification.* 243
Figure 5.13 | *Protein Identification*. 244
Figure 5.14 | *Protein Identification*. 246
Figure 5.15 | *Protein purification*. 247
Figure 5.16 | *Protein purification*. 248
Figure 5.17 | *Sequencing de novo*. 250
Figure 5.18 | *Sequencing de novo*. 251
Figure 5.19 | *Deglycosylation of pooled male urinary protein from T.vulpecula*. 252
Figure 5.20 | Development of a deglycosylation protocol for downstream mass spectrometry analysis. 254
Figure 5.21 | Development of a deglycosylation protocol for downstream digestion and sequencing. 256
Figure 5.22 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (2). 257
Figure 5.23 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (3). 258
Figure 5.24 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (4). 260
Figure 5.25 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (5). 261
Figure 5.26 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (6). 262
Figure 5.27 | *Structure of the tri-mannosyl core of* N-*glycans.* 264
Figure 5.28| *Deglycosylated sequencing.* 265
Figure 5.29 | *Protein sequencing*. 266
Figure 5.30 | *ESI-MS analysis of female brushtail possum urinary protein after purification and deglycosylation*. 268
Figure 5.31 | *Peptide sequence coverage from LC-MS/MS analysis of a purified 18 kDa protein from female brushtail possum urine*. 269
Figure 5.32 | Distinction of leucine and isoleucine using heavy isotope labelling. 271
Figure 5.33 | Distinction of leucine and isoleucine using heavy isotope labelling. 273
Figure 5.34 | *Residue conservation of the novel protein sequence .* 275
Figure 5.35 | *Phylogenetic analysis*. 276
Figure 5.36 | *Structural homology modelling*. 279

Figure 5.37 | *Comparison of protein identification results when searching peptide data from a global proteomics analysis of brushtail possum urine against a database comprising of all* Mammalia *sequences in SwissProt, or against all* Monodelphis domestica *sequences in UniProt.* 282
Figure 5.38 | *In-gel identification of brushtail possum urinary proteins*. 284
Figure 5.39 | *Label-free quantification*. 285
Figure 5.40 | *Label-free quantification*. 286
Figure 5.41 | *Label-free quantification*. 288

List of Tables

Table 1.1 | Examples of volatile semiochemicals reported to elicit a behavioural response in mammalian systems. 13

Table 2.1 | *ESI-MS/MS instrument settings.* 58

Table 3.1 | Cross-species matching with multiple species databases results in multiple identifications of the same protein. 101

Table 3.2 | Summary of protein identifications using cross-species matching against a multiple-species database and a single-species database..... 108

Table 3.3 | Identification results from LC-MS/MS analysis of bank vole bladder urine and scent marks. 130

Table 4.1 | *Possible mutation combinations of field vole glareosin.* 167

Table 4.2 | Top 20 most abundant peptides sequenced de novo in PEAKS™ from analysis of male scent marks. 208

Table 4.3 | Summary of novel and established lipocalin sequences and other notable protein identified in field vole urine and scent marks..... 215

Table 5.1 | *Protein identification.* 243

Table 5.2| Deglycosylation of pooled male urinary protein from T.vulpecula..... 253

Table 5.3 | *Structural Homology Modelling*. 278

Table 6.1 | Summary of chapter findings. 304

Supplementary Material

Supplementary material is attached.

S3 Characterisation of the protein content of scent secretions in the bank vole, Myodes glareolus
- S3.6 Supplementary Material
- S3.6.1 Published Paper
- S3.6.2 Individual Intact Mass Spectra
  - S3.6.2.1 Captive bank vole urine
  - S3.6.2.2 Wild bank vole urine
- S3.6.3 Glareosin sequencing fragment ion data (from paper supplementary)
- S3.6.4 Glareosin heavy leucine data
- S3.6.5 Multiple sequence alignment for phylogenetic analysis (from paper supplementary)
- S3.6.6 Bladder urine intact protein mass spectra
- S3.6.7 Female scent mark intact protein mass spectra

S4.6 Characterisation of the urinary protein content in the field vole, Microtus agrestis: Supplementary Material
- S4.6.1 Intact mass profiles of urine from mature field voles.
- S4.6.2 Fragment ion spectra of peptides used to sequence an initial draft of field vole glareosin.
  - S4.6.2.1 Initial sequence
  - S4.6.2.2 Mutations
  - S4.6.2.3 Peptide map of glareosin variants for individual males with distinctive protein profiles, used to confirm mutation sequencing.
- S4.6.2 Sequence coverage of odorant binding proteins from analysis of pooled female field vole urine.
- S4.6.2 Alignment of major urinary proteins to support preliminary sequencing of a field vole major urinary protein.
- S4.6.2 Intact mass profiles of juvenile field vole urine.

S5.6 Characterisation of the urinary protein content in the New Zealand brushtail possum, Trichosurus vulpecula: Supplementary Information
- S5.6.1 Intact mass profiles of individual samples
- S5.6.2 MALDI-ToF peptide mass fingerprints from tryptic in-gel digestion
- S5.6.3 Homologous sequences to the trichosurin-like protein from *M. domestica*
- S5.6.4 Fragment ion spectra for sequencing
- S5.6.5 Fragment ion spectra for determination of leucine and isoleucine residues
- S5.6.6 Table of accession numbers for multiple sequence alignment
- S5.6.7 Multiple sequence alignment of lipocalins

S6 Characterisation of urinary WFDC12 in small nocturnal basal primates, mouse lemurs (*Microcebus spp.*)

1 Introduction

Chemical communication between individuals in the natural world is integral to life on earth, and inter- and intra-species interaction is fundamental to survival of almost all; from simple quorum sensing in single cell organisms and fungal-plant symbiosis to establishment of complex hierarchical social structures in mammals.

Across the animal kingdom, semiochemistry is one of many tools extensively employed in the fight for survival. *Visual perceptions* are important in mate choice, evident in the iconic 'dance' of the male bird of paradise, displaying eye-catching plumage in the hopes of attracting a mate. *Tactile communication* is often used to improve the bond between individuals, as in the grooming behaviour of chimpanzees, and the importance of *auditory cues* is not only evident in the mating call of male frogs, but in the extensive use of language and inflection that we use as a species. Humans are comparatively anosmic compared to most of the animal kingdom, and may give less attention to the importance of chemical communication, defined as the conveyance of information through the use of, and subsequent detection of, chemicals, from one organism to another. Investigation into chemical signals has predominantly focused on insects, but there is a growing appreciation of the role that olfaction plays in non-human mammals.

1.1 Mammalian chemical scent communication

Chemical communication is reliant on the presence of a chemical cue, and the ability of a recipient to interpret it. Information can vary, from detection of a generic chemical profile, from which circumstantial information can often be acquired such as the health or sex of an individual, to signals that have evolved to convey a specific signal to a recipient. All chemicals involved in the interaction between organisms are termed *semiochemicals*, from the Greek for signal, which encompasses different groups. *Allelochemicals* are used to communicate information between individuals from different species, and can be divided further based on the nature of the communication. *Kairomones* are signals intended as a pheromone but are detectable to the benefit of an 'eavesdropping' species which can use those semiochemicals to serve their own purposes. *Allomones*, on the other hand, are examples of signals produced to mimic the semiochemical of another species to the detriment of the original signaller, whereas semiochemicals benefitting both species are termed *synomones*. A pheromone, however, is classified as a semiochemical used to convey information *within* a species.

Semiochemicals are involved in a wide variety of interactions within the animal kingdom, including territorial marking, parent-offspring interactions and mate choice/ reproductive success. For mammalian species, complexity of mediating chemical information is often reflective of complex social and environmental interactions. For example, an individual's chemical profile, composed of molecules directly from the individual in addition to the environment and external factors, could provide information on sex, species, reproductive status, health, social status and age. All this information is influential in natural selection, be that mate choice, offspring viability or kin survival, through interactions with other organisms. This could be with members of another species, or with conspecifics such as kin, competing individuals, offspring, or potential mates.

1.1.1 Social organisation

Social animals, particularly mammals and social insects, have complex societies. Maintenance of the social structures that are key to the survival of these species is dependent on interactions between individuals, which are often mediated by chemical cues. Often a learned response to a chemical profile, these can assist the animals in group recognition, promoting altruistic behaviours for kin survival.

Identification of individuals from the same colony or social group is evident in both social mammals and insects. For example, honeybees have guards to the hive, which prevent the

entry of parasites or conspecifics looking to steal food or larvae. This is achieved by the comparison of the chemical profile of the entrant to a 'template' profile, a learned odour profile representative of the hive (Getz, 1991; Couvillon *et al.*, 2007). In mammals, social groups are often comprised of family, and kin or group recognition can promote species survival, or kin reproduction, through co-operative behaviour. For example, juvenile spiny mice (*Acomys cahirinus*) learn the odour profiles of nestmates, irrespective of their relatedness, and treat unrelated nestmates the same as related nestmates later in life (Porter, Tepper and White, 1981). By contrast, ground squirrels (*Spermophilus beldingi*) respond to familiar, unrelated nestmates differently than to unfamiliar, unrelated individuals (Mateo, 2009). They later show less aggression to full siblings compared to half-siblings despite sharing the same nest (Holmes and Sherman, 1982), and to unfamiliar siblings compared to unfamiliar non-siblings, (Holmes, 1986), evidence of a genetic contribution to odour-mediated kin recognition. Similarly, responses to anal gland secretions of the beaver (*Castor canadensis*) elicited differing responses when extracted from an unfamiliar sibling compared to that of a non-sibling (Sun and Müller-Schwarze, 1997).

Kin recognition can also be used to avoid inbreeding. House mice (*Mus musculus domesticus*) are able to avoid inbreeding by detection of the major urinary proteins (MUPs), or their associated volatiles, a protein-mediated scent signal. The pattern in which these polymorphic gene variants are expressed is used, among other information, to identify kin and avoid inbreeding (Sherborne *et al.*, 2007). This is required in the house mouse due to the densely populated social structure.

Recognition of a conspecific as kin provides opportunity for investment in kin reproduction and the prevention of inbreeding. There is arguably a benefit to the inclusion of strangers into a social group for genetic variation, but with this the need to recognise an unrelated conspecific as a member of the group. Allomarking, marking of eachother, often by dominant members of the species, is exhibited in some mammalian social groups (Archie and Theis, 2011). For example, the European badger (*Meles meles*) scent mark each other with their subcaudal and anal glands (Kruuk, Gorman and Leitch, 1984).

Co-operative breeding in mammals is exhibited in a limited number of mammalian taxa, including rodents (*Rodentia*), dogs (*Canidae*), mongooses (*Herpestidae*), marmosets and tamarins (*Callitrichidae*). House mice suckle each other's young and co-operatively defend the nest, with no discrimination between pup relatedness (Lee, 2015). Common marmosets

(*C. jacchus*) exhibit singular co-operative breeding; only the dominant female breeds and the ovarian cycles of subordinate females are suppressed. A subordinate female can restart their ovarian cycle if isolated, but exposure to the scent from a dominant female can delay this by 20 days, but only if they are familiar; responding to the chemical profile, not a specific pheromone (Saltzman, Digby and Abbott, 2009). This is beneficial to the social group by reducing energy costs of a breeding pair.

1.1.2 Territorial marking

Many vertebrates use chemical signals to mark home ranges or territories, which can be used by conspecifics or by other animals to detect the presence of an individual or social group. This could be either a warning signal for other animals competing for resources or mates, a means of establishing dominance, an advertisement for potential mates, or a combination of the above, which could also be exploited by predators. In mammals, glandular secretions, urine or faeces are commonly placed in home ranges to convey the chemical signal. Generally, males tend to mark more frequently than females, particularly dominant males, which given the increased potential to hold resources, are often territory holders (Gosling & Roberts, 2001; Hurst & Beynon, 2004).

Different theories are propounded for the evolution of scent marking (Gosling and Roberts, 2001; Roberts, 2007; Wyatt, 2014). The scent-fence hypothesis proposes that the scent mark deters the entry of potential intruders into the territory. The scent-matching hypothesis, on the other hand, suggests it allows intruders to identify the owner of the scent mark, and therefore territory, upon meeting. Subordinate males can then choose to compete or withdraw based on scent mark and associated learning. Alternatively, the border-maintenance hypothesis theorises that scent marks are used to establish borders with major competitors.

Many group-living mammals use scent marks to defend the territory of the social group. For example, brown hyenas (*Hyaena brunea*) scent mark on top of those from other groups (Mills, Gorman and Mills, 1980), naked mole rats (*Heterocephalus glaber*) are able to recognise colony members and are aggressive to those outside the colony, even if closely related (O'Riain and Jarvis, 1997), and male ring-tailed lemurs flick testosterone-dependent secretions from wrist and shoulder glands towards opponents in defence of the social group (Scordato and Drea, 2007). Scent marking for territorial marking and scent marking for establishment of social hierarchies is not mutually exclusive. Dominant male and

breeding female house mice scent mark more frequently than other members of the group. Dominant males mark at particular signal sites, whereas subordinates mark at other consistent sites across the territory (Drickamer, 1995). Upon encountering another male, male mice are less likely to fight if their odour matches that of a previously identified scent mark in the territory, but this is dependent on competitive ability (Gosling and McKay, 1990). Dominant males are aggressive to subordinates that mark at the signal site, and counter-mark in response to an unfamiliar dominant male, but both of these responses are dependent on recognition of the individual odour as well as hierarchy of the individual (Hurst, 1993). They do not, however, counter-mark their own scent marks or those from a genetically identical male (Hurst, 1990).

1.1.3 Juvenile care

Semiochemical interaction also influences maternal behaviour, not only mediating maternal aggression (Wang and Storm, 2011), but also in identification of juveniles, maternal recognition and suckling behaviour (Nowak *et al.*, 2000; Schaal *et al.*, 2003; Arteaga *et al.*, 2013).

Olfaction is employed in recognition between mother and offspring. Beneficial to both, for the survival of offspring and the successful continuation of the mother's genetics, it is particularly important in mammals, as offspring are reliant on access to the mother's milk. For example, human babies are responsive to the odour of their mother (Schaal *et al.*, 2009), and ewes learn the odour of their lamb within the first few hours of the lamb's life (Lévy and Keller, 2008, 2009). An orphan lamb can be accepted during this time frame, but not after, suggesting that despite a genetic contribution to the lamb's odour (Porter *et al.*, 1991), it is not recognised by comparison to the ewes. These maternal behaviours and offspring recognition can be disrupted following induced anosmia to the main olfactory epithelium in primiparous ewes (Lévy *et al.*, 1995), although auditory and visual cues are still used (Ferreira *et al.*, 2000).

As mammalian young are dependent on milk, many species use chemical signalling to locate the mammary gland, attach to teat and suckle. The european rabbit (*Oryctolagus cuniculus*) nurse for only a few minutes every 24 hours, and mothers do not assist suckling by any tactile means (Hudson and Distel, 1983; Drummond *et al.*, 2000; Bautista *et al.*, 2005, 2008). The pheromone 2-methylbut-2-enal (2MB2) in milk is under hormonal control (Hudson, González-Mariscal and Beyer, 1990; Gonzalez-Mariscal, Chirino and Hudson, 1994)

and elicits a suckling response (Schaal *et al.*, 2003) but is not the only factor to cause the response. Domestic cats (*Felis silvestrus catus*) and dogs (*Canis lupus familiaris*) on the other hand nurse continuously for the first few days after the birth of a litter. Kittens and puppies exhibit nipple-searching behaviour, mediated by chemical cues under hormonal control (Arteaga *et al.*, 2013).

1.1.4 Sexual selection

The behaviour most strongly associated with olfactory signalling and pheromones is mating. The precedent was established by the discovery of bombykol, the female sex pheromone of the silk moth (*Bombyx mandarina*), in 1959 (Butenandt *et al.*, 1959), although the attraction response of the male moth to a hidden female, or the scent of a female, was first recorded as early as the 17th century by entomologist Jean-Henri Casimir Fabre (Fabre and Miall, 1921). It is used to attract the male over a distance of up to 10 km (Butenandt, Beckmann and Hecker, 1961). Since then, a wealth of pheromones has been identified that result in behaviours promoting reproductive success.

Intra-sexual selection is characterised by members of the same sex competing for a reproductive opportunity; this could involve competition for initial arrival to a potential mate, establishment of dominance over a competitor, or even post-copulatory sperm competition. Inter-sexual selection, however, is driven by mate choice based on the existence and prevalence of desirable phenotypes, or absence of those undesirable, including female sperm selection.

Considerably more effort has been dedicated to male signalling; many species exhibit sexual dimorphism with regards to scent mark behaviours, scent gland construction and chemical advertisement. Scent modulates male-male aggression in house mice; detection of male urinary cues by another male enhances aggressive behaviour towards the donor, but not towards female or castrated male intruders, unless anointed with non-castrated male urine (Mugford and Nowell, 1971; Mucignat-Caretta, Cavaggioni and Caretta, 2004). Male-male competition also extends beyond copulation; exposure to the urine of unfamiliar male house mouse will cause termination of a pregnancy in a recently mated female, named the ‘Bruce effect’ (Bruce, 1960).

In addition to male-male competition, male house mice also invest considerably in scent-mediated advertisement to the opposite sex. Mature females are more attracted to scent marks from sexually active males (Hurst, 1990; Petrulis, 2013) that deposit the greatest

number of scent marks, from which they can distinguish characteristics about the individual regarding male quality and competitive ability (Gosling and Roberts, 2001), preferring males that are dominant (Kruczek, 1997; Zhang, Zhang and Wang, 2001) and healthy (Kavaliers and Colwellt, 1995; Willis and Poulin, 2000), with high genetic quality (Thom *et al.*, 2008; Ilmonen *et al.*, 2009).

However, the wealth of information detected in a male house mouse scent mark can influence female response based on learned associations. Females prefer scents from male mice they have encountered, and develop a preference for volatiles present in a particular male's urine upon contacting that male's urine (Ramm *et al.*, 2008). However, this is modulated by the indication that the male is a territory owner (Gosling and Roberts, 2001).

Scent signals of the male house mouse can elicit a number of physiological responses in females: acceleration of the onset of puberty in juvenile females, named the 'Vandenbergh effect' (Nishimura, Utsumi, Yuhara, Fujitani, & Iritani, 1989; Novotny et al., 1999; Vandenbergh, 1969), initiation and synchronisation of oestrus in group-housed females (Koyama, 2004) and induction of ovulation (the 'Whitten effect') when females are exposed to the scent of reproductively active, dominant males) (Whitten, Bronson and Greenstein, 1968). The opposite effect, suppression of oestrus, occurs when female house mice are exposed to the urine of other female mice housed in single-sex groups (the 'Lee-Boot effect' (Van der Lee and Boot, 1956)), a situation which can delay puberty in juvenile female house mice (Colby and Vandenberg, 1974).

Despite a previous focus on male scent marking, females from many mammalian species exhibit scent marking behaviours with associated responses that promote reproductive success. Female scent marking increases in frequency during periods of receptivity and in the presence of male odours (Reviews: Coombes, Stockley, and Hurst 2018; R. P. Johnson 1973), which is suggested to increase competition in mates, therefore ensuring the female mates with the highest quality male (Rasmussen *et al.*, 1997). Some male mammals are able to distinguish between different stages in the oestrus cycle using odour cues, in which males are often more attracted to female scents during periods of receptivity (Johnson, 1973; Johnston, 1977; Birke, 1978; Smith, Mcdougal and Miquellet, 1989; Hudson and Vodermayer, 1992; Lai, Vasilieva and Johnston, 1996; Roberts and Dunbar, 2000; Ferkin, Lee and Leonard, 2004; Nie *et al.*, 2012). This may be particularly beneficial in solitary species, in which two members of the opposite sex are less likely to encounter each other

by chance, and olfactory-mediated advertisement is required to increase the chances of the interaction occurring (Wyatt, 2014).

In the house mouse, mature females scent mark in response to a male's scent mark to advertise sexual receptivity (Rich and Hurst, 1999), and males respond to female scent marks by producing ultrasonic vibrations, an important function of courtship behaviour (Nyby *et al.*, 1977). In terms of physiological response, scent marks of the female house mouse can cause increased sperm production in dominant males (Petrulis, 2013) and release luteinising hormone within half an hour of female urine exposure (Macrides, Bartke and Dalterio, 1975). Additionally, puberty in juvenile males is delayed if exposed to urine from females housed in single-sex groups (Jemiolo and Novotny, 1994).

Scent signalling of mammalian species is highly influential in mate choice. For females, a good mate selection is imperative considering the reproductive investment of females, and for males, olfaction provides another vehicle in which to advertise reproductive desirability (Stockley, Bottell and Hurst, 2013; Coombes, Stockley and Hurst, 2018).

1.2 The mammalian anatomy of pheromone detection

Pheromones are detected by the main olfactory and accessory olfactory systems, which ally distinct epithelial substructures and arrays of receptor molecules with discrete target regions in the olfactory bulb. These structures are conserved across the animal kingdom (Figure 1.1), reflecting the extent of which olfactory signalling and detection are used. Both systems are conserved within mammals, but vary in sensitivity and breadth of detection depending on the species.

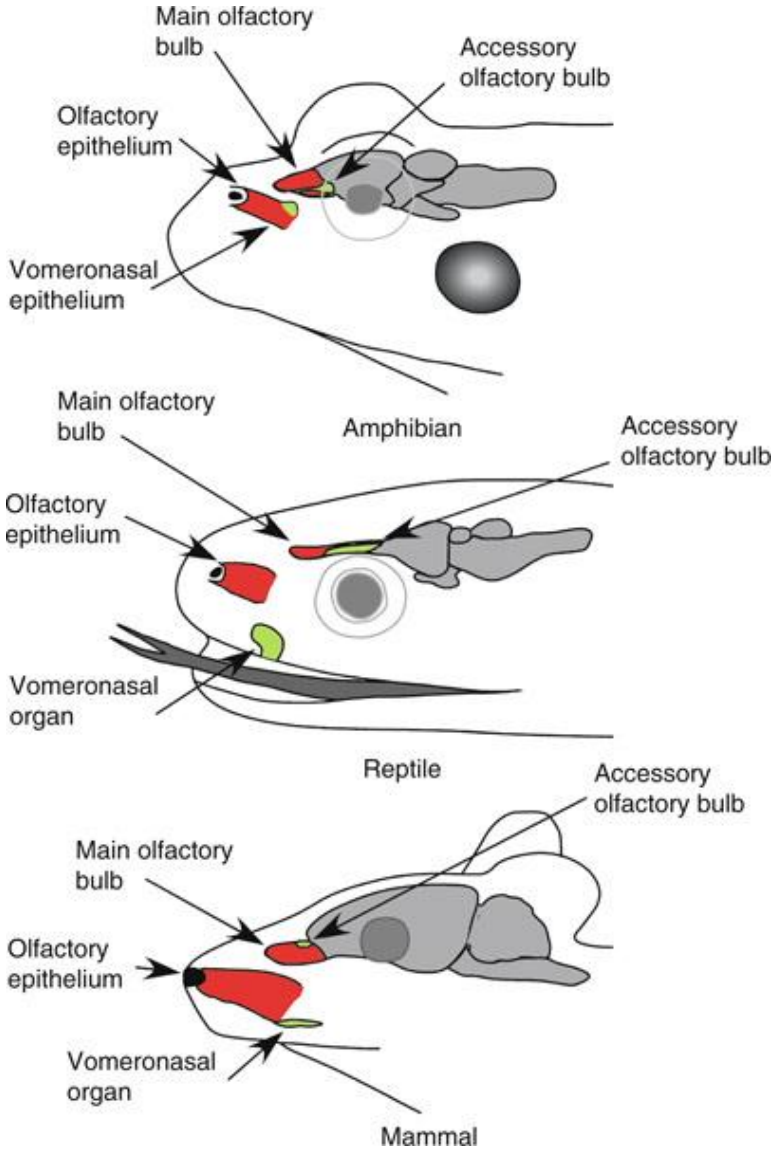


Figure from: (Martínez-Marcos and Halpern, 2009)

Figure 1.1 | *Structure of the nasal cavity across the animal kingdom.*
Many animals use two olfactory systems; the main olfactory system, in which the olfactory epithelium detects volatile signals, and the accessory olfactory system, in which the vomeronasal organ records involatile stimuli. Main olfactory system (MOE)

In the main olfactory system, volatiles are detected via G protein-coupled receptors (GPCRs) in olfactory sensory neurons located in the main olfactory epithelium (MOE) that detect predominantly airborne molecules. Two families of GPCRs are responsible for most chemoreception: the odorant receptors (ORs) and the trace amine-associated receptors (TAARs). ORs form a large multigene family, numbers of which vary from 400 in humans (Niimura, 2009) to 1576 in the rat and 1375 in the mouse (Zhang, Zhang and Firestein, 2007). The ligand-binding properties of individual ORs are overwhelmingly heterogeneous and promiscuous, capable of binding multiple odour molecules; similarly, single molecules can initiate signalling in multiple ORs (Munger, Leinders-Zufall and Zufall, 2009). Trace amine-associated receptors (TAARs) are a distinct, smaller class of GPCRs encoding 6 human genes and 15 mouse genes that preferentially invoke a response to amine compounds, among others (Liberles and Buck, 2006; Liberles, 2014). The MOE is instrumental in the detection of volatile mammalian pheromones, including 2-methylbut-2-enal, a lacrimal pheromone in the rabbit that induces suckling (Schaal *et al.*, 2003), and 2,5-dihydro-2,4,5-trimethylthiazole (TMT), a component of fox faeces which elicits an aversion response in rodents (Morrow *et al.*, 2000; Kobayakawa *et al.*, 2007).

1.2.1 Accessory olfactory system

In the accessory olfactory system, stimuli are detected by the vomeronasal organ (VNO), a fluid-filled sack that is instrumental in detecting non-volatile cues, and information is conveyed to the accessory olfactory bulb. The vomeronasal organ is strongly developed in snakes and lizards, and in many mammals such as rodents, marsupials and ungulates, but is argued to be non-functional in humans (Eisenberg and Kleiman, 1972; Liman and Innan, 2003; Zhang and Webb, 2003). As in the MOE, olfactory cues are detected by G-protein coupled receptors, of which there are three main expressed families: vomeronasal receptors types 1 and 2 (V1R and V2R, respectively) and the formyl peptide receptors (FPRs). Both types of vomeronasal receptor have large gene families in the rat: 115 intact V1R and at least 168 intact V2R genes are present, and in the mouse 191 intact V1R genes are present in contrast to 209 V2R genes (Yang *et al.*, 2005; Zhang, Zhang and Firestein, 2007). Distinct neuroepithelial layers make up the VNO; the layers behave as individual sub-systems within themselves, involving distinct classes of ligands, receptors and accessory olfactory bulb regions, both of which are capable of binding small volatile molecules in addition to soluble macromolecules through direct contact (Liberles, 2014). V1Rs are exclusively expressed in the apical zone and are known in mice to be responsive to steroidal

cues from urine, whereas V2Rs are expressed in the basal epithelium (Dulac and Axel, 1995; Meeks, Arnson and Holy, 2010; Isogai *et al.*, 2011), potentially function as heterodimers, and respond to protein and peptide ligands (Herrada and Dulac, 1997; Loconto *et al.*, 2003; Chamero *et al.*, 2007, 2011; Martini *et al.*, 2008; Leinders-Zufall *et al.*, 2009; Haga *et al.*, 2010). Expression of many V2Rs is correlated with expression of a subset of major histocompatibility complex (MHC) class I receptors, putatively proposed as chaperones (Ishii, Hirota and Mombaerts, 2003; Ishii and Mombaerts, 2008; Leinders-Zufall *et al.*, 2009). FPRs belong to a smaller gene family adjacent on the chromosome in mice to the vomeronasal receptor family, which has expanded in rodents, like most mammals, to function not only in the immune system, but also in the vomeronasal organ (Liberles, 2014). They are expressed in either the apical or basal layers depending on the type, but the functionality of these receptors is largely unknown.

The anatomy of the mammalian olfactory system varies between evolutionary lineages. The heterogeneous gene expansions of olfactory receptor types in the mouse and rat are indicative of an intensely sensitive detection of olfactory cues, capable of recognising and responding to a large range of ligands.

1.3 Chemical diversity of pheromones

An extensive range of chemical structures has been identified in mammals as semiochemicals, from small volatile amines to macromolecules such as the MUPs. Lower molecular weight semiochemicals, including both evolutionarily-determined pheromones as well as other chemical signals, are often products of metabolism, the microbiome or the result of hormonal changes. On the other hand, high molecular weight peptide and protein components are template-driven, directly encoded by inherited genes, therefore a direct display of genetic compatibility with the potential to influence mate choice. The chemical diversity of these molecules will be discussed in two sections: volatile, and involatile components.

1.3.1 Volatile semiochemicals

A wide range of volatile compounds cue a behavioural response in mammals (Table 1.1). These volatiles include ketones, acetates, terpenes, steroids, thiols, aldehydes, bicyclic acetals, amines, and pyrazines, and predominantly facilitate a response in odorant receptors of the main olfactory epithelium, although some stimulate receptors specific to the vomeronasal organ. The associated behaviours span a variety of interactions, and can influence not only sexual selection, but puberty acceleration or delay, pup suckling, predator avoidance, male-male aggression and maternal aggression. Published reviews (Apps, 2013; Liberles, 2014; Roberts *et al.*, 2014; Coombes, Stockley and Hurst, 2018) address the extent of mammalian chemosignalling and the pitfalls of mammalian semiochemical research. This includes exploration of signal complexity, and the likelihood that scent discrimination is dependent on the ratios of individual components, making up a unique combinatorial profile based on quantitative signals rather than simply qualitative. Some of the below pheromones are reported to work synergistically or only when in the presence of another component of urine, (Jemiolo *et al.*, 1985; Novotny *et al.*, 1985), which highlighted the importance of non-volatile components in mammalian odours and paved the way for characterisation of the major urinary proteins in the house mouse.

Table 1.1 | Examples of volatile semiochemicals reported to elicit a behavioural response in mammalian systems.
MOE = Main olfactory epithelium. VNO = vomeronasal organ. OR = odorant receptor. V1R = vomeronasal receptor type 1. V2R = vomeronasal receptor type 2. TAAR = trace amine-associated receptors.

Compound	Origin	Species	Receptor	Behavioural significance	Reference
(methylthio)-methylethiol (MTMT)	Male Urine	*Mus musculus* House mouse	OR (MOE)	Attraction of female	(Lin *et al.*, 2005)
(Z)-5-tetradecen-1-ol	Urine; preputial gland	*Mus musculus* House mouse	OR (MOE) Olfr288	Enhances urine attractiveness to females	(Yoshikawa *et al.*, 2013)
(Z)-7-dodecen-1-yl acetate	Female urine	*Elephas maximus* Asian elephant		Male flehmen response, erection, and premating behaviour	(Rasmussen *et al.*, 1996, 1997)
1-ido-2-methylundecane	Proestrus and oestrus female urine	*Mus musculus* House mouse		Investigation and attraction to urine by males	(Achiraman *et al.*, 2010)
2-*sec*-butyl-4,5-dihydrothiazole, dehydro-exo-brevicomin	Male Urine	*Mus musculus* House mouse	MOE V1R (VNO)	When both present; attraction of females, promotion of inter-male aggression, oestrus synchronisation and puberty acceleration.	(Jemiolo *et al.*, 1985; Novotny *et al.*, 1985; Leinders-Zufall *et al.*, 2000)
2,5-dihydro-2,4,5-trimethylthiazole(TMT)	Anal gland	*Vulpes vulpes* Fox	OR (MOE)	Fear response in rodents	(Kobayakawa *et al.*, 2007)
2,5-dimethylpyrazine	Adrenal glands of group housed females, secreted in urine	*Mus musculus* House mouse		Oestrus extension and puberty delay of both sexes	(Novotny *et al.*, 1986; Leinders-Zufall *et al.*, 2000)
2-heptanone	Male or female urine	*Mus musculus* House mouse	OR912-93 (MOE) V1RB2 (VNO)	Puberty delay	(Novotny *et al.*, 1986; Leinders-Zufall *et al.*, 2000; Boschat *et al.*, 2002; Gaillard *et al.*, 2002; Wang *et al.*, 2006)

2-methylbut-2-enal	Mammary	*Oryctolagus cuniculus* Rabbit	OR (MOE)	Suckling behaviour and odour induced learning in pups	(Schaal *et al.*, 2003)
2-phenylethylamine	Urine	*Lynx rufus* (bobcat) and other carnivores	TAAR4 (MOE)	Avoidance behaviour	(Liberles and Buck, 2006; Ferrero *et al.*, 2011)
6-hydroxy-6-methyl-3-heptanone	Male urine	*Mus musculus* House mouse	MOE V1R (VNO)	Puberty acceleration	(Leinders-Zufall *et al.*, 2000)
6-hydroxy-6-methyl-3-heptanone, n-pentylacetate, and isobutylamine		*Mus musculus* House mouse	VNO V1R	Maternal aggression & male sexual behaviour	(Del Punta *et al.*, 2002)
Androstenone (5α-androst-16-en-3-one)	Male saliva	*Sus scrofa* Pig	MOE	Positive reaction to lordosis in females	(Melrose, Reed and Patterson, 1971; Dorries, Adkins-Regan and Halpern, 1997)
Androstenone; androsta-4,16-dien-3-one	Sweat	*Homo sapiens* Humans	OR7D4 (MOE)	Variable perceptions of pleasantness (tentative)	(Keller *et al.*, 2007)
Cis-4-hydroxydodec-6-enoic acid lactone	Male tarsal scent	*Odocoileus hemionus columbianus* Black-tailed deer		Sex, familiarity and subspecies recognition	(Brownlee *et al.*, 1969; Muller-Schwarze *et al.*, 1976)
E,E-α-farnesene E-β-farnesene	Male preputial gland	*Mus musculus* House mouse	OSN20 (MOE) V1R (VNO)	Puberty acceleration	(Jemiolo, Xie and Novotny, 1991; Leinders-Zufall *et al.*, 2000; Wang *et al.*, 2006; Nara *et al.*, 2011)
Isoamylamine	Male urine	*Mus musculus* House mouse	TAAR3 MOE	Accelerates puberty onset in females	(Nishimura *et al.*, 1989; Liberles and Buck, 2006)
n-pentyl acetate *cis*-2-penten-l-yl acetate	Female urine	*Mus musculus* House mouse	MOE V1R (VNO)	Puberty delay	(Novotny *et al.*, 1986)
Trimethylamine	Male urine	*Mus musculus* House mouse	TAAR5 MOE	Attraction of female	(Liberles and Buck, 2006; Li *et al.*, 2013)

1.3.2 Involatile semiochemicals

Traditionally, semiochemicals are associated with volatile molecules. The precedent of this is from insects, whereby simple signals are released to stimulate aggregation of nearby conspecifics. However, involatile components are of equal importance in many mammalian species. Unlike volatiles, which are capable of attracting another animal to a scent mark, non-volatile cues are more stable and endure in the environment for a longer period of time, but require physical contact by the detecting animal. The vomeronasal organ, an architecture established to a varying extent within the animal kingdom, is responsible for the detection of water-soluble, macromolecular cues. For example, sulfated steroids in the urine of the female house mouse are detected with class specificity by V1R receptors (Nodari *et al.*, 2008; Isogai *et al.*, 2011). However, many of the macromolecules influential in semiochemical function are peptides or small proteins.

Proteins are essential in the production and detection of all scent marks: either through enzymatic action for production or metabolism of the detected components, by providing specialised detection of the scent molecules in the olfactory structures, or by direct action as a component in the deposited scent mark. The protein complexity of deposited scent marks can vary considerably between species; house mice express high quantities of polymorphic protein species into their urine, resulting in a complex, sexually dimorphic, information-rich scent mark (Robertson *et al.*, 1996; Beynon and Hurst, 2003; Sherborne *et al.*, 2007), whereas the Roborovskii hamster excretes a single predominant protein species in the urine of both sexes (Turton *et al.*, 2010). The mammalian scent mark proteome appears so far to be highly species-specific in terms of complexity and function, although expansive research is required, particularly beyond the well-characterised scent signalling of the house mouse, to understand the roles of proteins in scent communication across *Mammalia*. The following section discusses the function and expression of proteins previously established in mammalian scent marks.

1.4 Mammalian Proteins in Scent-Mediated Communication

Whilst few proteins have been directly implicated in a chemosignalling role at present, many species exhibit high levels of protein in secretions commonly used for scent marking. Protein production is metabolically expensive, and external secretion of significant quantities of protein is indicative of an important function, however proof of a semiochemical or even pheromonal function requires evidence that the odorant consistently elicits a specific response, determined by behavioural assays.

In humans, proteinuria is a marker of glomerular filtration malfunction; urine produced in a healthy human kidney contains only trace amounts of protein when corrected for urine dilution (0.08 mg/mL (Newman *et al.*, 2000)). In comparison, in healthy house mice, urinary protein output is typically 10-15 mg/mL, and is sexually dimorphic; wild female house mice excrete approximately 30% less than males (Beynon and Hurst, 2003; Armstrong *et al.*, 2005). Scent marking secretions in mammals are often rich in a single type of protein, ranging from multiple variants of the same type (such as MUP expression in the house mouse) to a single predominant protein species. This is evident in the high abundance of urinary protein cauxin, a carboxylesterase in feline species involved in the production of 2-amino-7-hydroxy-5,5-dimethyl-4-thiaheptanoic acid (felinine) (Miyazaki *et al.*, 2003, 2008; McLean *et al.*, 2007), and the whey acidic protein, WAP four-disulfide core domain protein 12 (WFD12) expressed seasonally in the urine of Madagascan mouse lemurs (Unsworth *et al.*, 2017).

Cauxin and WFD12 are both examples of enzymes, changing the chemical profile of the scent mark whilst deposited. Cauxin is excreted into urine, and is an esterase in the pathway that produces felid-specific felinine. Felinine decomposes to a number of species-specific compounds, including 3-mercapto-3-methyl-1-butanol (MMB) which, in faeces, has been identified as a male sex recognition pheromone (Miyazaki *et al.*, 2018). WFD12 on the other hand contains a whey acidic protein motif, which in some cases have established antimicrobial properties (Hagiwara *et al.*, 2003; Yenugu *et al.*, 2004; Wilkinson *et al.*, 2011). Whilst the role of WFD12 present in Madagascan lemur urine is as yet unknown, antimicrobial activity would also influence the chemical profile of the urine (Unsworth *et al.*, 2017). Both enzymes have the potential to orchestrate chemical signals within respective scent secretions, and the metabolic expense of excreting large quantities of protein suggests this is a highly likely explanation for their presence.

Scent signalling proteins can also function as ligand carriers. In particular, the barrel-shaped MUPs in the house mouse are capable of binding small, volatile, active molecules within the central calyx of the protein, for they exhibit ligand-binding specificity (Darwish Marie *et al.*, 2001; Phelan *et al.*, 2014). MUPs promote the slow release of associated volatiles, extending the lifespan of the scent mark (Hurst *et al.*, 1998), so that unlike auditory and visual communication, the signal is sustained in the environment after the signaller has left. Bound ligands include the semiochemically-active pheromones that suppress ovulation in females and accelerate puberty in juvenile females (Novotny *et al.*, 1986; Jemiolo and Novotny, 1994). MUPs can also act as pheromones in their own right through direct detection by V2R receptors (Chamero *et al.*, 2007; Roberts *et al.*, 2012; Kaur *et al.*, 2014).

A range of non-volatile species is suggested to have roles in scent signalling. Many belong to the lipocalin superfamily, including not only MUPs, but also OBPs and probasins, which are discussed in section 1.4.4. Other non-volatile species include androgen-binding proteins, exocrine gland-secreting peptides, and major histocompatibility peptides.

1.4.1 Salivary androgen-binding protein (ABP)

Androgen-binding proteins (ABPs) are glycosylated members of the secretoglobin family that are synthesized in the Sertoli cells of the testis in many mammalian species and are capable of binding testosterone (Hagenäs *et al.*, 1975). ABPs are the products of a rapid gene expansion and evolution in the house mouse, in which a salivary ABP has been identified (Laukaitis *et al.*, 2005). The α-subunit, which forms a heterodimer with either the β or γ-subunit, has undergone a micro-evolution in house mouse sub-species (Hwang *et al.*, 1997). Female members of *Mus musculus domesticus* and *M. m. musculus* discriminate towards territories of, and show mate preference for, males based solely on the ABPα from submaxillary and lacrimal gland expression (Laukaitis, Critser and Karn, 1997).

1.4.2 Exocrine gland-secreting peptides (ESPs)

Exocrine gland-secreting peptides (ESPs) are the result of another gene cluster, some products of which are also secreted via the extraorbital lacrimal gland of the house mouse. House mice and Norway rats (*Rattus norvegicus*) both have multigene clusters, containing 24 and 10 ESP genes respectively (house mice also have 14 pseudogenes), which is in

contrast to humans, who have none (Kimoto *et al.*, 2007). The mice and rat multigene clusters encode a set of peptides of approximately 7 – 10 kDa, mainly secreted into the tear duct via the extraorbital lacrimal gland, but some are also secreted by the harderian or submaxillary glands. Expression of two ESPs in the extraorbital lacrimal gland of BALB/c mice is sexually dimorphic; ESP36 is more highly expressed in females, and ESP1 is male specific, although some variation, particularly for ESP36, is observed in different mouse strains (Kimoto *et al.*, 2007). Expression is hormonally controlled; castrated males do not express ESP1, but express ESP36, but this is reversed following testosterone propionate treatment. Likewise, testosterone propionate-treated females exhibit suppressed expression of ESP36 but induced expression of ESP1 (Kimoto *et al.*, 2007). Male ESP1 secretion in tear fluid enhances female receptivity to lordosis behaviours by stimulating the vomeronasal receptor V2Rp5 in females (Kimoto *et al.*, 2005; Haga *et al.*, 2010) and is influential in the chemosensory-stimulated pregnancy termination (the 'Bruce effect'); female mice exhibit high rates of pregnancy failure after encountering males with a different quantities of ESP1 compared to the female, but this effect was not observed in *V2Rp5*$^{-/-}$ females (Hattori *et al.*, 2017). Exposure to ESP1 is also thought to enhance male-male aggression (Hattori *et al.*, 2016).

1.4.3 Major Histocompatibility Complex (MHC) Peptides

The MHC consists of highly polymorphic loci, whose genes encode cell surface glycoproteins for immune recognition of pathogens. Its main function is considered to be immunological, encoding proteins with structurally diverse regions that bind and display pathogenic peptide antigens for host recognition at the cell surface (Novotny *et al.*, 2007). When antigens are released from the MHC molecule outside of the cell, they are excreted in urine and other secretions, and the structures of these antigen molecules, often peptides or oligosaccharides, which mirror the antigen-binding site of MHC molecules, provides a molecular identity signal (Singh, Brown and Roser, 1987).

The heterogeneity of the MHC locus has been proposed to assist a number of different social interactions, although the molecular mechanism by which this occurs is not yet established. MHC molecules have been proposed as possible chaperones to V2Rs (Ishii, Hirota and Mombaerts, 2003; Loconto *et al.*, 2003; Ishii and Mombaerts, 2008), and synthetic MHC peptide antigens stimulate signalling in olfactory sensory neurons (Spehr *et al.*, 2006) and cause a broad but specific response in the vomeronasal organ to structurally

different peptides (V2R1b) (Leinders-Zufall *et al.*, 2009). However, the extent to which signalling occurs at natural ligand concentrations is unknown (Leinders-Zufall *et al.*, 2004; Chamero *et al.*, 2011), although an influence of the MHC on mate choice, inbreeding avoidance, parent-offspring interactions, and pregnancy block has been suggested (Yamazaki *et al.*, 1983, 2000; Boyse, Beauchamp and Yamazaki, 1987; Brennan, 2009).

Laboratory strains of the house mouse show preference for mates with a dissimilar MHC type (Yamazaki *et al.*, 1976; Penn and Potts, 1998), and it is hypothesised that this would increase MHC heterozygosity of potential offspring, increasing resistance to disease. However, there is no evidence to support immunological advantage (Ilmonen *et al.*, 2007). It has also been linked to kinship recognition and inbreeding avoidance; house mice learn the MHC profiles of related individuals to discriminate kin and prevent inbreeding, but this can be reversed by cross-fostering (Penn and Potts, 1998). The MHC complex has, however been linked to the 'Bruce effect'; exposure of pregnant mice to MHC peptides from a different mouse strain results in pregnancy failure (Leinders-Zufall *et al.*, 2004). However, MHC-associated odours are not sufficient, nor necessary, for scent owner recognition (Hurst *et al.*, 2005). Therefore, whilst MHC type can be discriminated by odour recognition in the house mouse (Penn and Potts, 1998), there is no evidence so far to suggest that MHC molecules are independently influential in individual recognition.

1.4.4 Lipocalins

1.4.4.1 Structure and Function of Lipocalin Proteins

Lipocalins make up a large protein family that has a high level of functional diversity, both within species and between. They are characterised by their highly conserved tertiary structure, an eight-stranded antiparallel β-barrell architecture with ligand-binding affinity for predominantly hydrophobic molecules (Figure 1.2) (Flower, 1996). Each beta-sheet is interspersed with loops (L1-7) typical of short β-hairpins, except the larger loop 1 (L1), which forms a lid. Also characteristic of its tertiary structure is an N-terminal 3_{10} α-helix (A1), and a longer C-terminal α-helix (A2) (Flower, 1996). However one of the most prominent aspects of lipocalin architecture is the ligand-binding calyx, from which it derives its name, and which has shown to have a strong affinity for a wide range of low molecular weight ligands.

Whilst the physical structure of lipocalins is highly conserved, there is high levels of disparity between sequences; sequence conservation is often below 20%. Flower (1996)

classifies lipocalins into two groups (Figure 1.3); the first, kernel lipocalins, contain three structurally conserved regions. They include retinoic acid-binding protein-4, apolipoprotein D, complement C8 gamma chain, prostaglandin D2 synthase and major urinary proteins. The second group is comprised of outlier lipocalins, which have lower levels of sequence similarity and may only contain one or two of these motifs. These include odorant binding proteins, von Ebner's gland protein and probasin. The defining sequence motif is the tripeptide G-X-W at the N-terminal, which is common to all lipocalins. The second is a region within the sixth and seventh beta-sheets, and the third is a conserved arginine residue in the eighth beta-sheet (Figure 1.4). All three are highlighted on a multiple sequence alignment of representative, relevant lipocalin sequences (Figure 1.4). Ganfornina *et al*. (2000) notes that whilst lipocalins share an affinity for binding hydrophobic ligands, the capacity of the internal calyx tends to be smaller in outlier lipocalins such as MUPs and OBPs than kernel lipocalins.

The MUPs are kernel lipocalins, and show concordance with all three structurally conserved motifs. Examples are shown in Figure 1.4; mouse MUP1 (the solved structure of which is displayed in Figure 1.5) and darcin, to represent central and peripheral mouse MUPs, respectively (UniProt Accessions P11588 and Q5FW60), and a rat MUP (UniProt Accession P02761). OBPs and probasins on the other hand are outlier lipocalins, lacking evidence of both C-terminal motifs (Figure 1.4). Interestingly trichosurin, lacks the N-terminal G-X-W motif thought to be common to all lipocalins. However X-ray crystallography has shown that trichosurin maintains the conserved tertiary structure of lipocalins, albeit in an unusual homodimeric arrangement (Figure 1.6) (Watson *et al.*, 2007), and contains the conserved arginine residue towards the C-terminal. Phylogenetic analysis of lipocalin sequences indicates distinct classes, which may assist in attributing putative functions to novel sequences based on homology (Figure 1.7). It also allows an assessment of lipocalin evolution; *Mup* gene expansion in the Norway rat and the house mouse is extensive, compared to limited polymorphism in OBP genes of each species.

Whilst the overall structure of lipocalins is conserved (Figure 1.5, Figure 1.6), the binding capabilities can vary considerably depending on the primary sequence, changes of which can alter surface chemistry, both externally and within the central cavity. Post-translational modifications also alter the external chemistry of lipocalins; incidences of glycosylation are frequent, and the site is not conserved between those established sites (Figure 1.5; Figure 1.6). Due to the potential pheromone-binding capabilities of externally secreted lipocalins, sequence variation has a putative role in olfactory communication.

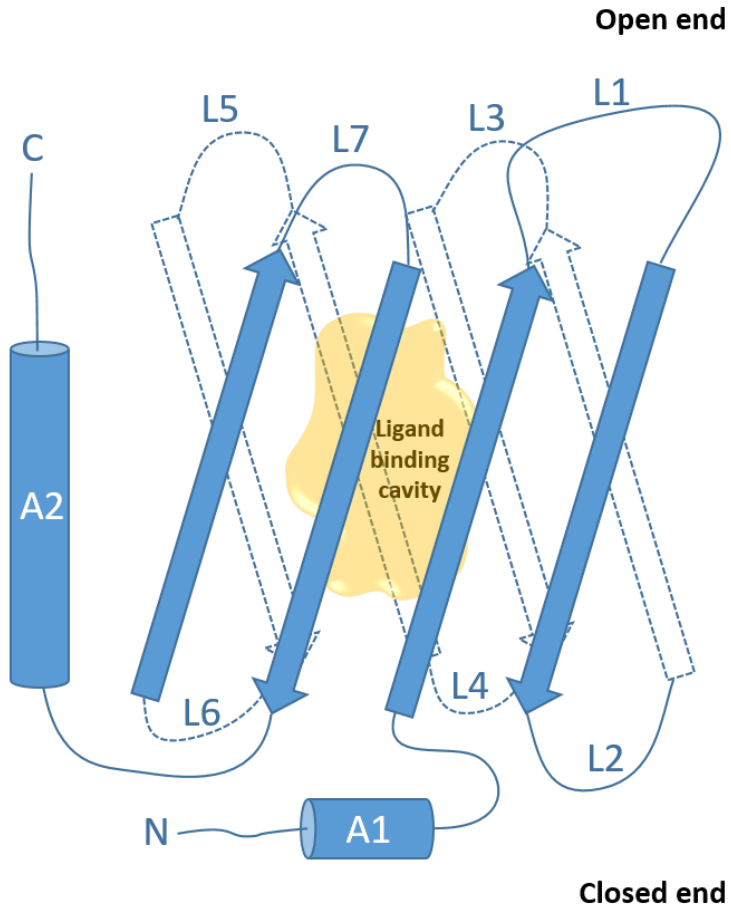


Figure 1.2 | *Structure of lipocalins.*
Redrawn from (Ganfornina *et al.*, 2000). Lipocalin structure consists of eight antiparallel beta-sheets interspersed with β-hairpin loops (L1-7), whereby L1 is larger than the rest and forms a 'lid' over the 'open end' of the barrel structure. Other features include an N-terminal and a C-terminal alpha-helices, and an internal cavity capable of binding small, largely hydrophobic ligands.

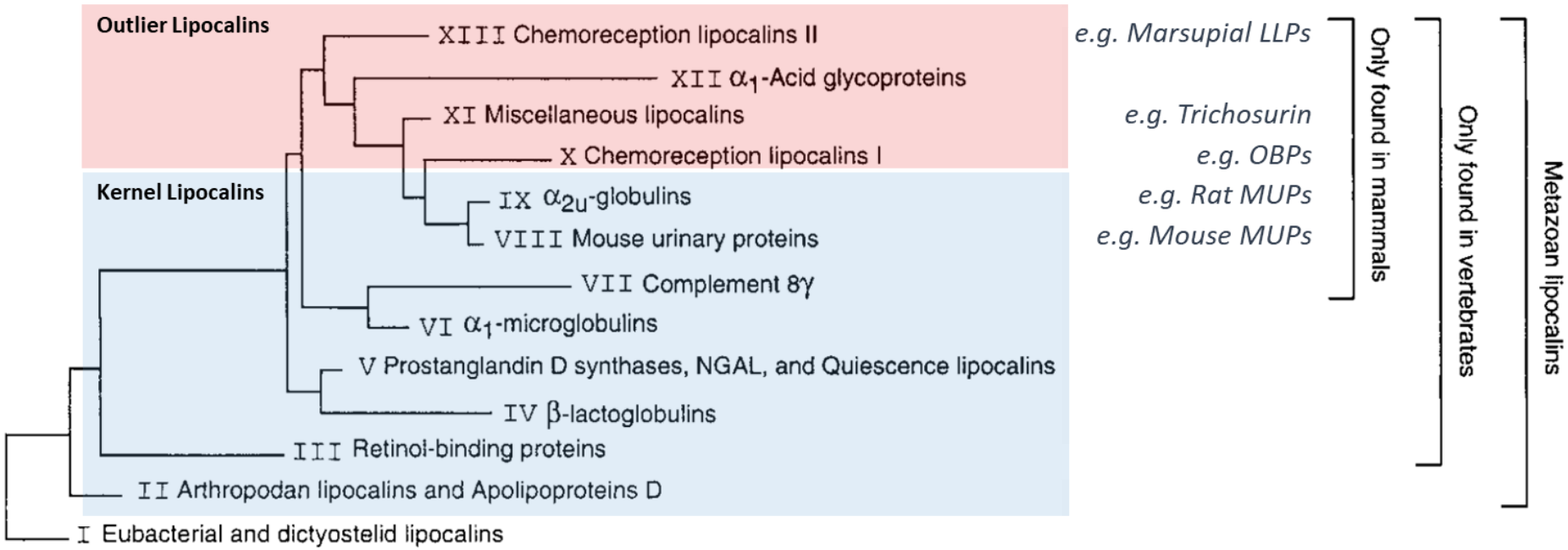


Figure 1.3 | *Classes of lipocalins.*
Annotated, from (Ganfornina *et al.*, 2000). Phylogenetic tree of metazoan lipocalin clades, labelled as kernel and outlier lipocalins according to Flower (1996).

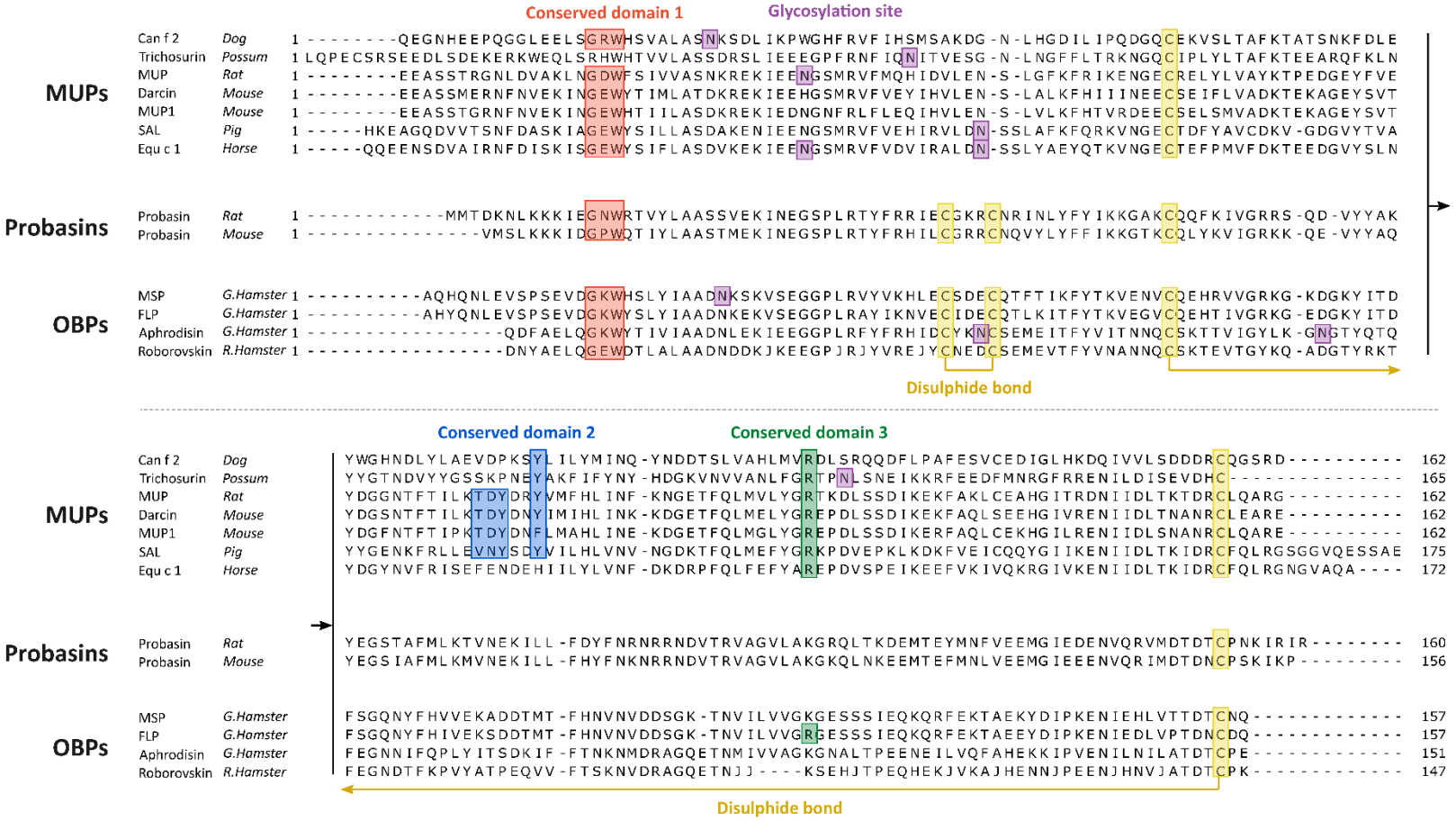


Figure 1.4 | *Sequence conservation in the lipocalin family.*
Conserved sequence domains characteristic of lipocalin sequences are indicated on representatives of MUP sequences (kernel lipocalins), probasins (outlier lipocalins) and odorant binding proteins (outlier lipocalins). Signal peptides were removed (Petersen *et al.*, 2011), and a multiple sequence alignment was generated in Clustal Omega (Sievers *et al.*, 2011).

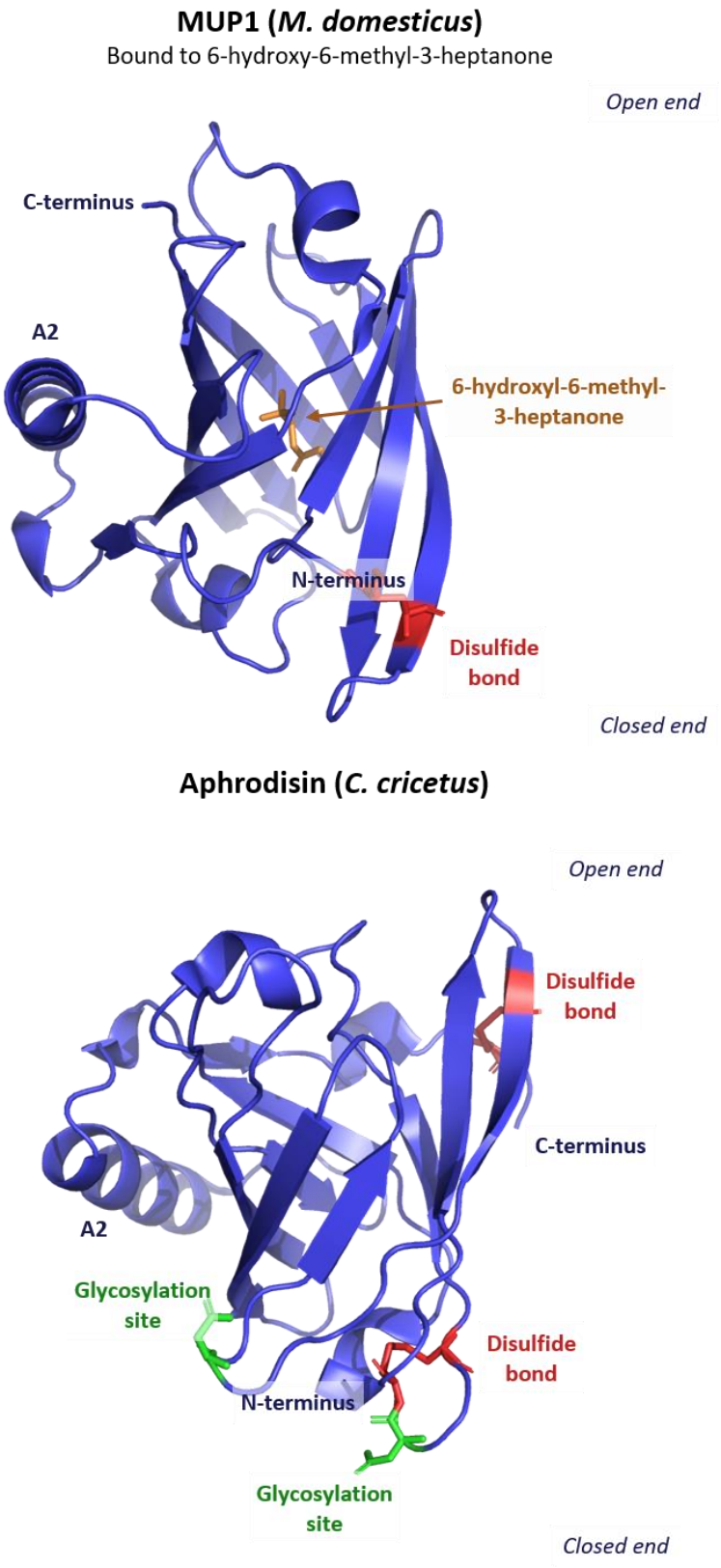


Figure 1.5 | *Structure of lipocalins.*
Top; structure of MUP-1 crystallised with 6-hydroxyl-6-methyl-3-heptanone (*M. musculus*) PDB 1I05 (Timm *et al.*, 2001). Bottom; Aphrodisin (*C. cricetus*), PBD 1E5P (Vincent *et al.*, 2001). Structures were downloaded from RCSB Protein Data Bank (https://www.rcsb.org/), and formatted in PyMOL (PyMOL, no date).

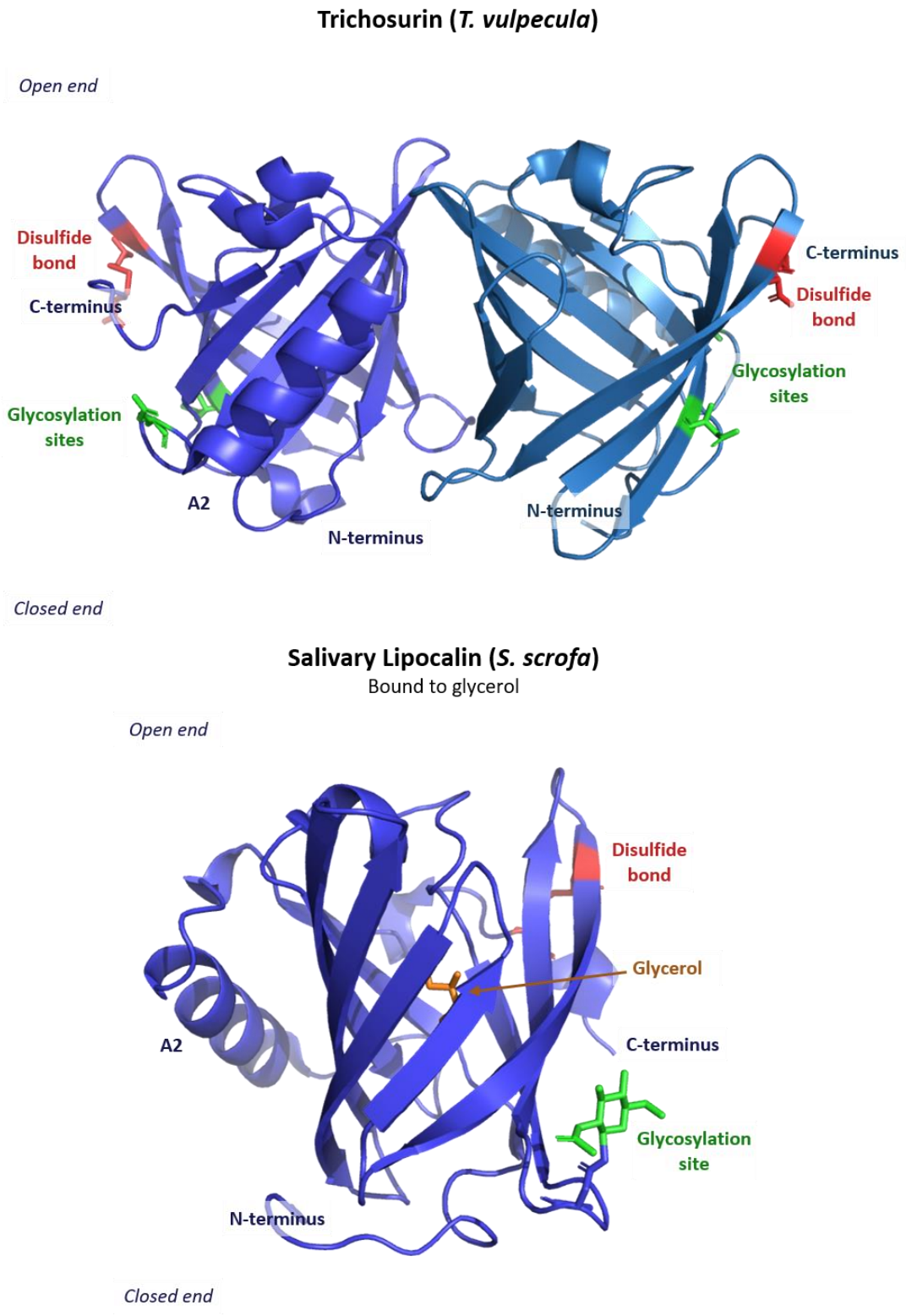


Figure 1.6 | *Structure of lipocalins.*
Top; structure of Trichosurin (*T. vulpecula*) PDB 27R3 (Watson *et al.*, 2007). Homodimer chains are coloured differently. Bottom; Salivary lipocalin solved bound with glycerol (*S. scrofa*), PDB (Spinelli *et al.*, 2002). Structures were downloaded from RCSB Protein Data Bank (https://www.rcsb.org/), and formatted in PyMOL.

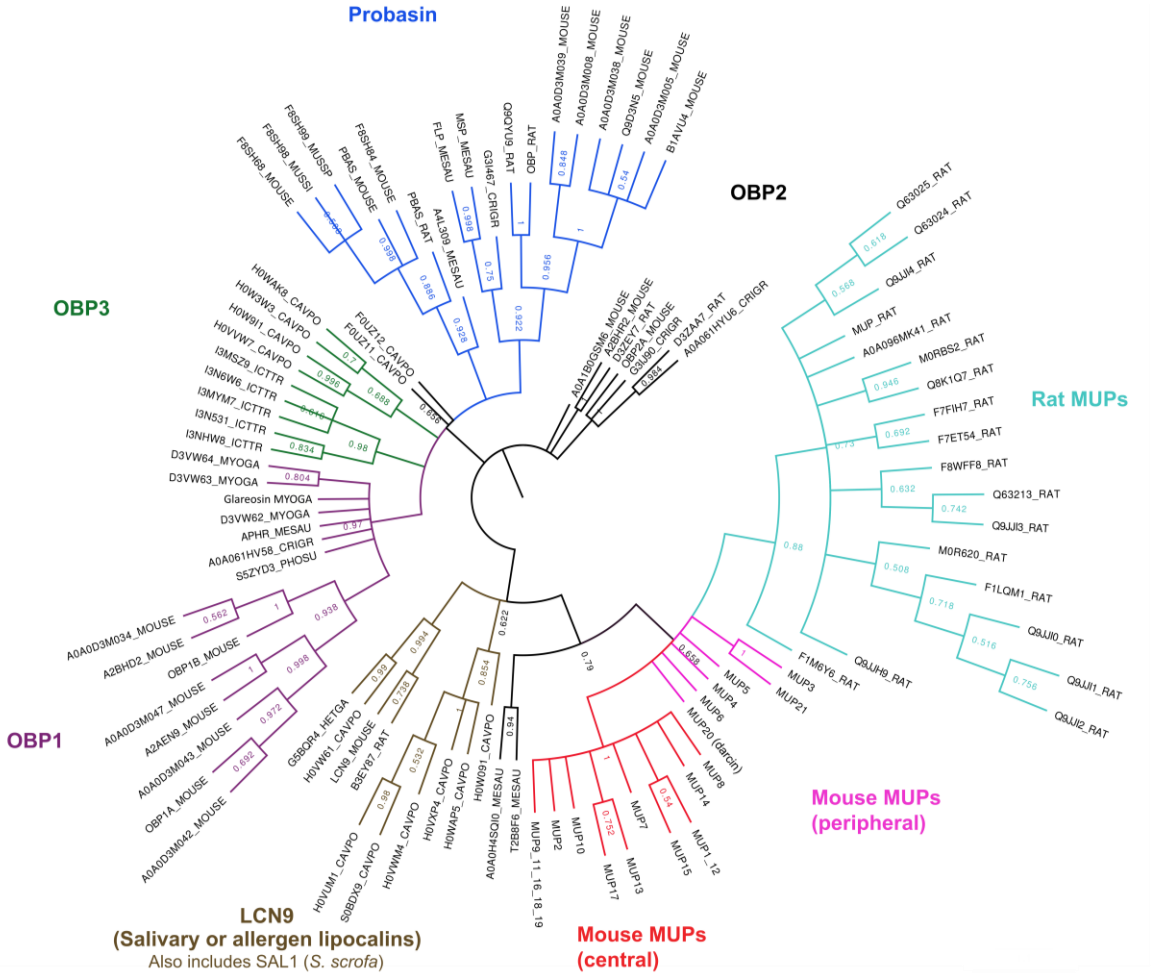


Figure 1.7 | *Phylogeny of lipocalin sequences in rodents.*
Phylogenetic analysis of sequences from lipocalin sub-classes in rodents with links to scent communication. Annotated, from Loxley *et al*. (2017).

1.4.4.2 Major Urinary Proteins (MUPs)

The best studied signalling proteins are the major urinary proteins (MUPs), discovered in the urine of the house mouse (Finlayson and Baumann, 1957; Finlayson *et al.*, 1965). These liver-expressed (Hastie and Held, 1978) proteins have arisen from rapid gene expansion, resulting in highly polymorphic urinary protein profiles (Knopf, Gallagher and Held, 1983). Based on the C57BL/6 genome sequence (GRCm38), there are a minimum of 21 genes encoding MUPs in the house mouse, located in a cluster on chromosome four (Mudge *et al.*, 2008). Of these, 15 are classed as central MUPs and show high levels of sequence similarity (97%), likely resulting from more recent gene duplication events than peripheral MUPs (82-94% similarity) (Mudge *et al.*, 2008). They have a typically conserved structure of the lipocalin family (Flower, 1996), an eight-stranded β-barrel structure with a central hydrophobic calyx capable of binding apolar ligands (Bacchini, Gaetani and Cavaggioni, 1992)

Whilst MUPs are present in both males and females, sexual dimorphism is pronounced, with an overall MUP output three- to four- times higher in males, and some variants are unique to males (Hurst and Beynon, 2013). Phenotypic polymorphism results from the stable expression of a MUP profile that differs between individuals, particularly in wild mice (Hurst *et al.*, 2001; Beynon *et al.*, 2002; Hurst and Beynon, 2004). MUPs are instrumental in establishing dominance, mate attraction and recognition of kinship and individual identity (Hurst *et al.*, 2001; Beynon *et al.*, 2002; Chamero *et al.*, 2007; Cheetham *et al.*, 2007; Michael Garratt *et al.*, 2011; Nelson *et al.*, 2015; Lopes and König, 2016; Sheehan *et al.*, 2016).

One function is to promote the lifespan and slow release of associated volatile ligands (Hurst *et al.*, 1998; Beynon and Hurst, 2004; M. Garratt *et al.*, 2011) thereby extending the life of the scent mark. MUPs are capable of binding low molecular weight hydrophobic compounds, including known pheromones 2-*sec*-butyl-4,5-dihydrothiazole and 3,4-dehydro-*exo*-brevicomin (Bacchini, Gaetani and Cavaggioni, 1992; Robertson, Beynon and Evershed, 1993; Novotny *et al.*, 1999). One male-specific peripheral MUP, darcin, has highly specific ligand-binding properties, binding the male-specific pheromone 2-*sec*-butyl-4,5-dihydrothiazole and promoting its steady release over many hours, itself being unusually stable (Armstrong *et al.*, 2005; Phelan *et al.*, 2014). It also directly stimulates vomeronasal 2 receptors (Chamero *et al.*, 2007), essential for a learned attraction to the associated volatile cues, and therefore the individual male that deposited the scent mark (Roberts *et al.*, 2012).

MUPs are also expressed in the urine of the Norway rat, *Rattus norvegicus*. Rat MUPs, previously known as α_{2u}-globulins, are expressed to a lesser degree of polymorphism than in the mouse, but with stronger sexual dimorphism (Vandoren *et al.*, 1983; Gómez-Baena *et al.*, 2014). Rat MUPs are expressed in other tissues, including salivary, lachrymal, Meibomian, mammary, preputial and perianal glands, whereas no evidence so far suggests preputial gland MUP expression in mice (Held and Gallagher, 1985; Gómez-Baena *et al.*, 2014). The role of rat MUPs in terms of semiochemical properties is not yet understood, but urine is utilised in conveyance of sex, age, reproductive status, hierarchy and individuality (Brown, 1988, 1995; Gómez-Baena *et al.*, 2019).

Previous genome analysis of 13 vertebrates, 12 of which were mammals and included the mouse and rat MUP genes discussed above, suggested many mammalian species, including the pig, dog, chimpanzee, orangutan, macaque, bushbaby and opossum, have only one functional protein-coding *Mup* gene (Logan, Marton and Stowers, 2008). This is except for the horse which has three, and the mouse lemur which has two (Logan, Marton and Stowers, 2008). These are more evolutionarily distant than mouse and rat MUPs, and many are denoted in the literature as allergens or salivary lipocalins (Figure 1.7).

The protein encoded from the pig *Mup* gene is expressed in boar saliva, called salivary lipocalin (SAL) (Booth & White, 1988; Marchese *et al*., 1998). Its expression is sexually dimorphic, with no evidence of the protein in the submaxillary gland of the sow, and two isoforms have been isolated differing by three amino acids (Loebel *et al.*, 2000). It is capable of binding the boar pheromones androstenone and androstanol, the binding of which increases the volume of the internal cavity to accommodate the associated steroids (Spinelli *et al.*, 2002). Both forms are glycosylated in the same position and the tertiary structure is held by a single disulfide bridge (Loebel *et al.*, 2000; Spinelli *et al.*, 2002). The same proteins are also expressed in the mucosa of both sexes, however the attached glycan residues differ (Marchese *et al.*, 1998).

The *Mup* gene in the dog, Can f 2, is also expressed in the saliva. It has been identified as an allergen with another lipocalin (Can f 1) that is more closely related to the von Ebner's gland protein, a lipocalin known to be expressed in lingual salivary glands with a possible role in taste perception (de Groot *et al.*, 1991; Bläker *et al.*, 1993). Can f 2 contains a potential glycosylation site in the same position when aligned with the pig salivary lipocalin (Figure 1.4). One of the three horse *Mup* genes has been identified as major horse allergen, Equ c 1. It is expressed in sublingual glands, submaxillary glands and liver, and is

glycosylated on two sites (Gregoire *et al.*, 1996) (Figure 1.4). Whilst a product of the opossum *Mup* gene has not been identified at the protein level so far, another lipocalin has been identified with homology to MUPs in another marsupial, the common brushtail possum, *Trichosurus vulpecula* (Piotte *et al.*, 1998). Trichosurin is expressed in the mammary gland with two other lipocalin-like proteins (β-lactoglobulins), is capable of binding small phenolic compounds, and exists as a dimer (Watson *et al.*, 2007). Watson *et al*. (2007) hypothesised that the ligand-binding properties of the protein indicate a possible role in priming the liver of the neonate to produce the enzymes required to metabolize otherwise toxic plant phenols, however no potential function is thus far supported with any evidence.

1.4.4.3 Odorant Binding Proteins (OBPs)

OBPs are expressed in nasal tissue in a wide range of species (Bignetti et al. 1985; Briand et al. 2000; Dal Monte et al. 1991; Lazar et al. 2002; Pes et al. 1992; Pes and Pelosi 1995) and are thought to facilitate the transport of small signalling molecules across the mucosal membrane. However, they are also expressed in urine, vaginal secretions, saliva, seminal fluid and lacrimal glands (Briand, Trotier and Pernollet, 2004; Rajkumar *et al.*, 2010; Stopková *et al.*, 2010a; Turton *et al.*, 2010; Dubey *et al.*, 2013; Mastrogiacomo *et al.*, 2014; Loxley *et al.*, 2017).

In mice, a lower level of OBP polymorphism is observed at the genome level in relation to MUPs, occurring in a cluster of six intact genes on the X chromosome (Stopkova *et al.*, 2009; Stopková *et al.*, 2014). Whilst a deficit of genome data for other species prevents an assessment of genome-level polymorphism, protein-level polymorphism has been observed to some degree in the nasal mucosa of the porcupine (*Hystrix cristata*) (Felicioli *et al.*, 1993), nasal tissue and saliva of the giant panda (*Ailuropoda melanoleuca*) (Zhu *et al.*, 2017) and the urine of the bank vole (*Myodes glareolus*) (Stopková *et al.*, 2010a). Other secretions reported thus far exhibit single, abundant proteins.

Aphrodisin is a small, 17 kDa glycosylated lipocalin that is expressed in the vaginal secretion of female golden hamsters (*Mesocricetus auratus*) and is the best characterised signalling protein in muroid rodents (Briand, Trotier, & Pernollet, 2004; Singer & Macrides, 1990). Both vaginal marking behaviours and aphrodisin expression varies across the reproductive cycle, with the former peaking prior to receptivity and the latter peaking at oestrus, mediated by levels of estradiol and progesterone (Johnston, 1977; Lisk and Nachtigall, 1988; Briand, Trotier and Pernollet, 2004). An OBP, it is a member of the lipocalin family and shares the same conserved lipocalin structure with internal ligand-binding calyx as MUPs (Henzel *et al.*, 1988; Vincent *et al.*, 2001). Aphrodisin binds small hydrophobic ligands, and the combined effect of aphrodisin and low molecular weight ligands promotes mounting behaviour in males (Singer *et al.*, 1986; Loïc Briand *et al.*, 2000; Briand *et al.*, 2004).

In the Roborovski hamster, *Phodopus roborovskii*, high levels of the lipocalin roborovskin is seen in the urine of both sexes, lacking both the extreme polymorphism and sexual dimorphism observed in rats and mice (Turton *et al*., 2010). Male and female members of *Mesocricetus auratus*, the Syrian hamster, both exhibit sexual dimorphism. In the submandibular salivary gland of the male, a glycosylated OBP, male-specific secretory

protein (MSP), is expressed at high levels (Dubey *et al.*, 2013). In the extraorbital lacrimal glands, females of the same species produce a female-specific extraorbital lacrimal gland protein (FLP) (Ranganathan and De, 1995). Both are hormonally controlled and share 85% similarity, and are 58% identical to rat OBP (Ranganathan, Jana and De, 1999).

In the seminal fluid of the rabbit (*Oryctolagus cuniculus*), an OBP3-like protein was identified, with sexual dimorphism in terms of expression within the reproductive systems (Mastrogiacomo *et al*., 2014). An OBP has also been identified in buffalo saliva (Rajkumar *et al*., 2010).

Probasins are androgen-regulated outlier lipocalins that, based on phylogenetic analyses, are closely related to other identified OBPs (Ganfornina *et al.*, 2000; Loxley *et al.*, 2017). These proteins of approximately 19 kDa are distinguished in the literature due to their specific expression in the prostate, as reported in mice and rats (Matuo *et al.*, 1982, 1984; Johnson *et al.*, 2000; Kasper and Matusik, 2000). They have an unusually high pI of 11.5, and a larger proportion of intron in its DNA sequence; whilst a similar molecular weight protein product to other lipocalins, the gene size is approximately 6 times greater (Kasper and Matusik, 2000). Probasin expression in the rat has been proven to be androgen-regulated by castration and subsequent androgen treatment studies, with levels increasing upon sexual maturation (Kasper and Matusik, 2000).

1.5 Methodology of scent mark analysis

Identification of a semiochemical involves multidisciplinary steps. Either the location of, expression of, or behaviour associated with a secretion may suggest olfactory function. Consequent examination of its chemical composition may produce a list of candidates whose characteristics are worthy of further investigation; perhaps a subset of molecules are only expressed in the breeding season, or in the presence of a conspecific. Bioassays testing the behavioural or physiological response of the putative active compound are then required for validation.

The ability to chemically characterise scent marks has improved both in throughput and sensitivity in the past decades due to the advance of analytical techniques. Approaches for the analyses of scent marks are multidisciplinary, but primarily involve mass spectrometry.

In mammals, active semiochemicals are typically embedded in a complex sample, such as glandular secretions, urine or faeces, from which the active component, or components, must be extracted. There is also considerable discussion regarding the combinatorial effects of scent mark components. Nevertheless, determining the pheromone, or pheromones, that elicit a biological response to a scent mark is reliant on careful investigation of not only the molecular entity as a whole, but essentially, the individual components. The approach taken will vary with sample complexity, volatility, molecular weight, polarity and available instrumentation.

Approaches used for the extraction of volatile components of scent marks include solvent-based extraction, solid-phase microextraction (SPME) and stir bar sorptive extraction, which isolate the volatile constituents from the remainder of scent marks. The volatiles are usually separated using gas chromatographic techniques for identification using an allied mass spectrometer (GC-MS) (Zhang, Yang and Pawliszyn, 1994; Soini *et al.*, 2005). For elucidation of the associated behaviours, identified compounds are often synthetically reproduced and employed at biological concentrations in behavioural assays (Novotny, 2003).

Analysis of non-volatile components requires alternative methods. Lipids have been extracted from scent marks for downstream analysis following chloroform/methanol separation followed by GC-MS (Bligh and Dyer, 1959; Poddar-Sarkar, 1996; Foster, 2005). Assessment of protein content on the other hand, is predominantly reliant on separation techniques allied with mass spectrometry, such as liquid chromatography- mass spectrometry (LC-MS) or liquid chromatography-tandem mass spectrometry (LC-MS/MS).

Proteomics has proven instrumental in the identification, characterisation and quantification of scent mark proteins, particularly with the improvement of range, scope and depth of analytical capabilities in recent years.

The advance of instrumentation has permitted an increasingly complete analysis of protein content. Proteins in a biological sample may have a large dynamic range of abundances, potentially compromising proteomic analysis by interference of dominant proteins. Additionally, proteome complexity extends beyond that of the genome; it is multi-dimensional, accounting for gene splicing events, isoforms and post-translational modifications. A systematic approach to sample complexity is required to tease apart different aspects of the proteome in order to then identify and quantify proteins. In terms of scent marking, this may be used toreveal differences between samples from males and females, dominant and subordinate individuals, or intact and castrated males, for example.

1.5.1 Assessment of sample complexity

Initial assessment of sample complexity is required to plan more in-depth subsequent investigation. Polyacrylamide gel electrophoresis (PAGE) (Davis, 2006) has been used extensively to assess protein complexity, separating according to electrophoretic mobility and staining to detect resolved proteins. When PAGE is performed in native form, it can provide information on the conformation and charge of molecules relative to one another, in combination with the molecular weight. More commonly however, proteins are denatured using sodium dodecyl sulfate to disrupt their tertiary and quaternary structures and given a uniformally distributed negative charge so that separation is based almost entirely on molecular weight alone, a variation called SDS-PAGE (Laemmli, 1970). This technique has been used extensively in the early identification of proteinuria in mammals and is a fundamental biochemical technique for assessing sample protein complexity; the number of resolved bands is indicative of the different protein species, and the intensity and width of staining is proportional to the protein abundance (Gordon *et al.*, 1993; Beynon *et al.*, 2013). 2D-PAGE separates proteins in two dimensions by combining these techniques, first by their isoelectric point, then their molecular weight in the perpendicular direction, and has been instrumental as a separation technique. However, with the vast improvements in on- and off-line separation techniques prior to mass spectrometric analysis, 2D-PAGE has become used far less frequently than one dimensional electrophoresis, which is predominantly used as a first-pass assessment of protein complexity as it is less labour- and resource-intensive.

An additional assessment of protein complexity is mass spectrometric analysis of the intact proteins within a sample. Development of ionisation techniques electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI) eliminated the restriction of protein upper molecular weight limits that previous ionisation techniques weren't capable of accessing (Ho *et al.*, 2003; Wilm, 2011). Both are soft ionisation techniques that cause little fragmentation of the ion, meaning that mass-to-charge ratios of intact proteins can be determined when coupled with a mass spectrometer. ESI is capable of producing multiply charged ions, which form an envelope of multiply charged peaks in the mass spectra and can be deconvoluted to determin the overall mass of the protein species (Ho *et al.*, 2003).

With a single protein, intact protein analysis can be utilised to gather information on protein folding, proteoforms and post-translational modifications (PTMs). It can give an indication of protein complexity at a higher resolution than compared to PAGE, providing the level of complexity does not prevent distinction of individual protein envelopes. For example, the similarity of MUPs means that the extensive polymorphism is not observed to its full extent by either SDS or native PAGE, however intact mass analysis by ESI-MS has been capable of distinguishing between the isoforms (Armstrong *et al.*, 2005).

Intact protein mass profiles have been generated to investigate MUP variation between the sexes, for example (Armstrong *et al.*, 2005). Whilst intact mass can provide a wealth of information regarding proteome complexity and post-translational modifications, it leaves much information undefined at the sequence level. Top-down proteomics, where intact proteins are fragmented to gather sequence-level information, is still a developing field and most protein identification is reliant on the common bottom-up proteomics approach.

1.5.2 Protein identification

Before the advance of high-throughput mass spectrometry capabilities in the application of proteomics, proteins were identified in low complexity or isolated samples by the generation of a peptide mass fingerprint (PMF). A protein, or a small number of proteins, is digested with a site-specific protease, cleaving the amino acid sequence at certain sites to generate peptides. The *m/z* of these peptides is then measured by MS and a profile of peptide masses is generated as a unique 'fingerprint' of this protein. When comparing to a database of proteins digested theoretically at the same cleavage site, an identification of the protein can be made based on the combination of peptide fragments (Perkins *et al.*, 1999). This technique has been frequently employed after separation of proteins by SDS-, native or 2D-PAGE to identify individual proteins within a sample. However, PMF

identification is only effective for low complexity samples, and consequently relies upon prior separative techniques, in addition to database matching of peptides. PMF is therefore an unsuitable option for identification in complex samples, however, it can still be a useful technique for fast identification of low complexity samples.

The advance of instrumentation and software has improved both identification and quantification of proteins,, and the selectivity, speed, and precision with which this is performed. Bottom-up proteomics using tandem MS is by far the most common approach taken for high-throughput protein identification in complex samples; proteins are digested with a site-specific protease, usually trypsin, and resulting peptides are isolated for further fragmentation within the mass spectrometer to generate both precursor ion *m/z* values and their corresponding fragment ion spectra. These precursor ion values are most commonly searched against a database of theoretically digested proteins to produce a list of candidate peptides with corresponding precursor masses, from which theoretical fragment ion spectra are compared with spectra in the experimental data to identify peptide-spectrum matches (PSMs), thus identifying the corresponding proteins (Cottrell, 2011). Search engines such as MASCOT (Perkins *et al.*, 1999) and SEQUEST (Eng, McCormack and Yates, 1994) facilitate database identification, and are frequently used with publicly-available databases such as UniProtKB (Apweiler *et al.*, 2004) or NCBI RefSeq (Pruitt, Tatusova and Maglott, 2005), resources which provide a comprehensive, searchable knowledge base of proteins, compiled from genome, transcriptomics and literature-derived data. User-defined parameters are adjusted to the query. For example, tolerance of the precursor and fragment ion masses can be set according to the mass analyser: orbitrap mass analysers are capable of much higher mass accuracy, so tolerances can be set to lower values in comparison to ion traps, for example. Search-defined modifications are set to increase the number of peptide-spectrum matches by allowing for mass increments corresponding to known modifications on a given residue. This may be allowing for modifications that occur during sample preparation, or in search of biologically relevant PTMs. Increasing the number of variable modifications, however, increases the search space exponentially, so searches need to be balanced with computational power available. Some software algorithms employ multiple-round searching to search multiple PTMs whilst maintaining reasonable search times, by increasing the parameters to include many variable PTM possibilities, but restricting the search space in other ways. For example, PEAKS™ performs an optional secondary search of up to 650 variable PTMs in an algorithm called PEAKS PTM™ , but only maps unidentified spectra sequenced *de novo* by the

software with a high quality cut-off score to those proteins already identified from the first-round search (Han *et al.*, 2011). Other open modification search tools include SeMoP (Baumgartner *et al.*, 2008), TwinPeaks (Havilio and Wool, 2007) and MSFragger (Kong *et al.*, 2017).

Search algorithms also commonly incorporate a validation method for large datasets called a decoy search, whereby a decoy database, for example of the reversed protein sequences in the search database, is searched (Elias *et al.*, 2005). Matches from the decoy database are termed false positives, and the rate of false positive matches divided by the total number of positive matches is given as the false discovery rate (FDR). It is therefore common to adjust the significance threshold of results to give the probability of falsely identifying a protein 1%.

However, this approach is reliant on the availability of a database of relevance to the experiment in question. Whilst the genomes of model species such as the house mouse are comprehensively sequenced and extensively annotated, protein identification in other species is often challenging. Cross-species matching employs peptide-spectrum matching of experimental data to the available proteomes of closely-related species, with some degree of success. However, this approach is most effective with closely conserved proteins such as housekeeping proteins. For proteins under higher evolutionary pressure, such as those with roles in sexual selection and mate competition, protein-coding genes experience higher rates of evolution (Ramm *et al.*, 2008; Karn and Laukaitis, 2009; Bayram *et al.*, 2016). Sequence mutations reduce the number of viable peptide-spectrum matches with which to identify proteins, with the potential to omit identification altogether, or reduce confidence of the identification to that below the typical parameters used. For example, the number of unique peptides for confident identification is often set to a minimum of three as standard within the field. This presents an additional issue when using a bottom-up approach to investigate gene products from gene duplication events, such as MUPs. MUPs are so similar in sequence that identification using a bottom-up proteomics approach is reliant on very few unique peptides capable of distinguishing between MUP variants.

PEAKS™ software employs an additional optional search algorithm, SPIDER™, that detects single point mutations. PEAKS™ sequences *de novo* all spectra, giving an overall and local confidence scores, and uses these sequences to perform a database search (Ma *et al.*, 2003). It then employs an optional multi-round strategy, performing an open modification search (PEAKS PTM) and mutation search (SPIDER) on the unidentified sequenced spectra,

mapping modified sequences to proteins identified in the first database search. The SPIDER search takes into account common sequencing errors, non-specific digestion and mass changes associated with single point mutations to detect peptides homologous to those in the database (Han, Ma and Zhang, 2005). However, it is limited firstly by requiring an initial identification of the protein from the first database search, and secondly by requiring a user-defined number of matched amino acids, both of which are reliant on a reasonably high level of homology in the database protein. For housekeeping proteins, the number of mutations is not expected to be high, whereas for proteins under high evolutionary pressure, even a SPIDER search may not provide an identification.

Consequently, a careful approach, is required when using cross-species matching. For novel proteins with putative biological relevance in species without an available reference genome, isolation and sequencing *de novo* offers a more thorough, albeit more labour-intensive, methodology.

1.5.3 Sequencing de novo

Comprehensive sequencing and annotation of genomes other than those of model organisms would require a research effort not achievable in the present day. We are therefore faced with the challenge of analysing functionally relevant proteins without a reference database. Full characterisation of proteins requires complete sequence information, which, following isolation of the protein, is achievable through sequencing *de novo*.

Common protein purification techniques take advantage of differences in molecular weight, charge, hydrophobicity, affinity or biological activity, and are dependent not only on the protein in question, but also on the residual biological matrix. Isolation of MUPs, for example, has been successful following the application of ion exchange chromatography; the high similarity in the expressed gene products renders separation using size exclusion techniques near impossible (Robertson *et al.*, 1996, 2007; Armstrong *et al.*, 2005; Cheetham *et al.*, 2009).

Protein sequencing *de novo* is reliant on the ability of a mass spectrometer to fragment peptides and obtain fragment ion spectra. The fragmentation method employed is paramount to deciphering the spectra, as different ions can be produced using different methods. Collision-induced dissociation (CID), or higher-energy collisional dissociation (HCD) in Orbitrap instruments, uses collisions with a neutral gas to fragment peptides, obtaining predominantly b- and y-ions (Figure 1.8). These ions occur from the breakage of

the peptide backbone, so that a complete series of b- or y-ions would be incremental by the mass of the next amino acid residue in the peptide sequence. The y-ion series is generally more stable than the b-ion series, particularly in HCD fragmentation methods, as b-ions are less likely to survive the additional fragmentation (Medzihradszky and Chalkley, 2015). Additionally, the b1 ion is rarely observed; b-ions are formed from a nucleophilic attack of the neighbouring N-terminal carboxyl group, which doesn't exist for the b1 ion (Medzihradszky and Chalkley, 2015)

Mass increments within a b- or y-ion series can estimate a peptide sequence for each fragment ion spectra, but relies on good quality spectrum. Figure 1.9 shows two examples of fragment ion spectra; one of them has an easily defined fragment ion series from which the sequence could be determined (A), and one shows poor spectral quality (B).

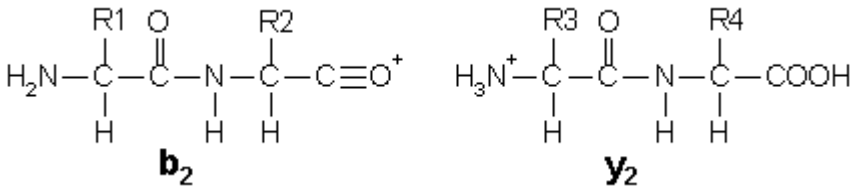


Figure 1.8 | *Structure of b- and y-ions.*
Figure from MASCOT (http://www.matrixscience.com/help/fragmentation_help.html.)

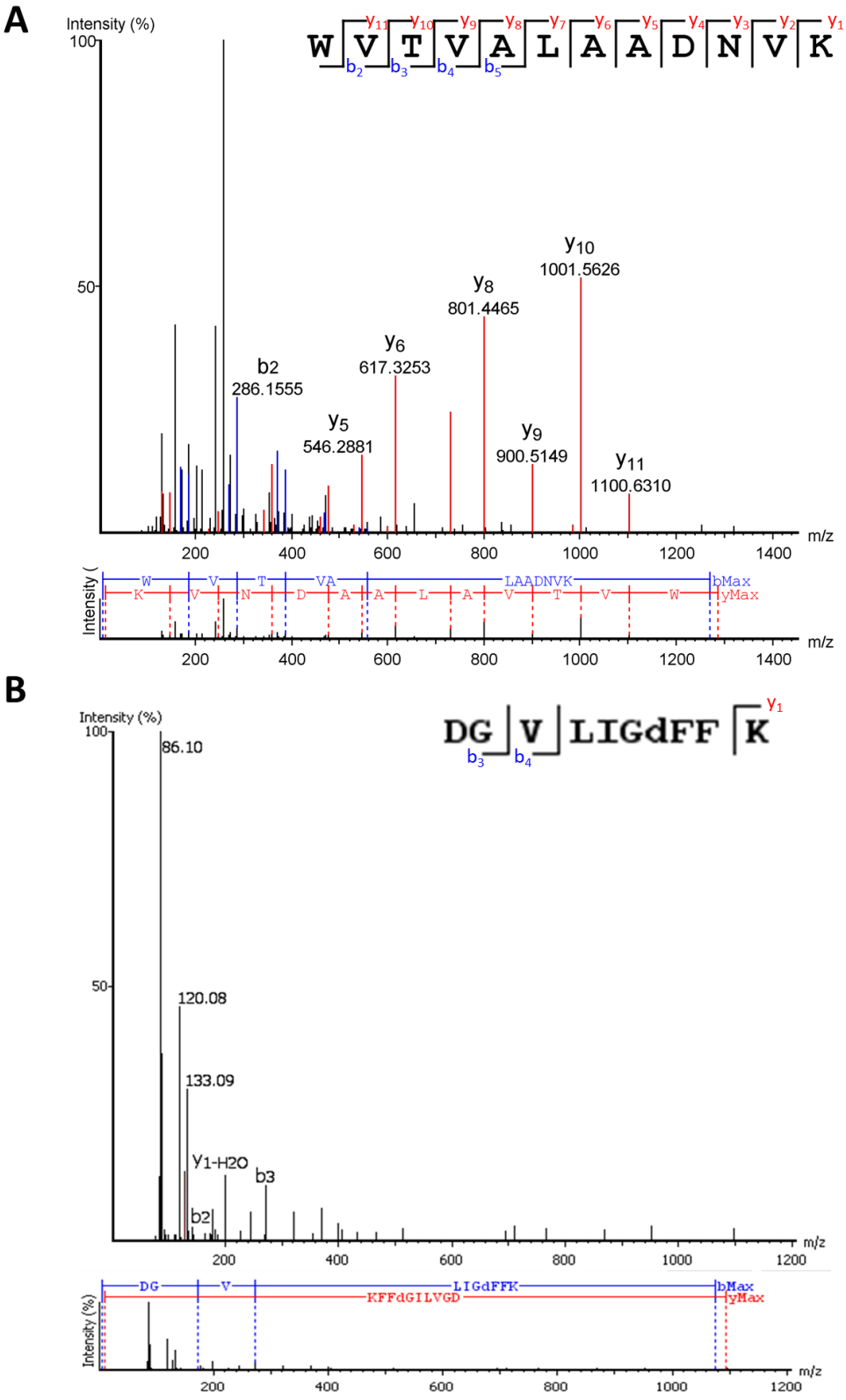


Figure 1.9 | *Example fragment ion spectra.*
Fragment ion spectrum displaying identified b- and y-ions from the sequencing *de novo* of urinary proteins from male field voles and female brushtail possums. A; example of a high-quality PSM, B; example of a low-quality PSM.

Whilst sequencing *de novo* can be labour-intensive to do manually, programmes have been developed to automatically sequence spectra *de novo*. In the 1980s, an algorithm was developed to produce all possible amino acid combinations from spectra and use the resulting candidates for database searching (Sakurai *et al.*, 1984), a technique still employed; programmes such as pNovo, PepNovo, and PEAKS™ (Ma *et al.*, 2003; Frank and Pevzner, 2005; Chi *et al.*, 2010) employ algorithms that suggest peptide candidates sequenced *de novo* for a given spectrum, and use these generated sequences to provide an additional level of support to techniques in database matching for protein identification, like those techniques discussed in section 1.5.2.

Furthermore, some *de novo* sequencing algorithms employ sequence tagging methods, developed in the 1990s (Mann and Wilm, 1994), that search well-characterised segments of the fragment ion spectra. This not only improves database search time for identification but is also used to match homologous peptides by searching accurately sequenced portions of peptides (Chi *et al.*, 2010). For example, PEAKS™ software produces '*de novo* tags' that partially match a sequenced fragment ion spectrum to a database sequence. These *de novo* tags are based on matching a user-defined number of consecutive amino acids (default is five), even if the remaining predicted amino acids of the sequenced spectrum or the precursor mass do not match.

The sequencing of peptides *de novo* is the first stage of protein sequence construction. Sequenced peptides can either be aligned against a scaffold of homologous proteins, or overlapping peptides can be generated by digestion of the protein with multiple proteases with complementary site-specific cleavages, and aligned forthwith, the chosen approach for the sequencing of Roborovskin, as demonstrated in Figure 1.10 (Turton *et al.*, 2010).

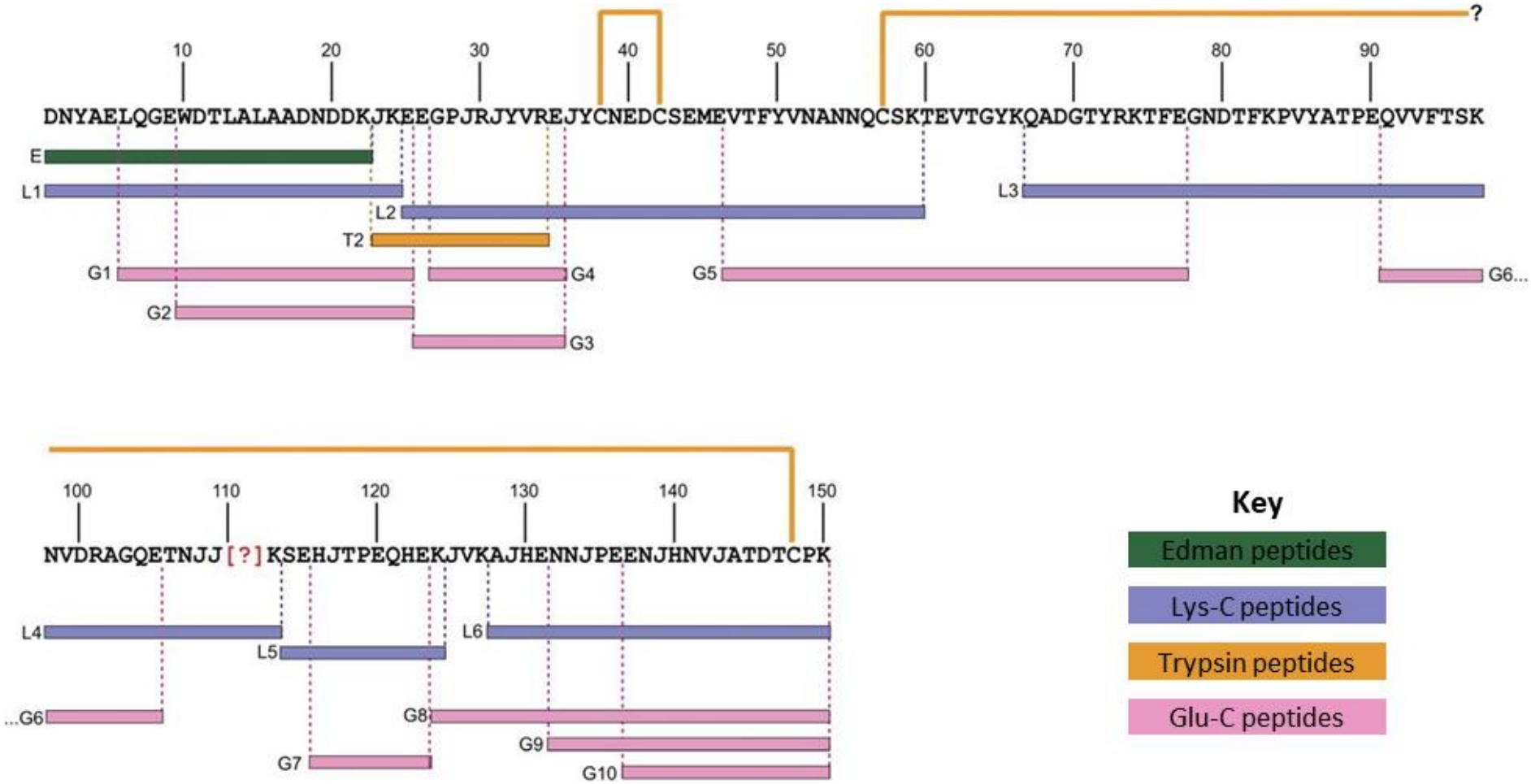


Figure 1.10 | *Sequencing* de novo *using overlapping peptides.*
Annotated, from Turton *et al*. (2010). Roborovskin, a protein in the urine of the *Roborovskii* hamster, was sequenced *de novo* using overlapping peptides generated from multiple specific proteases.

One issue with protein sequencing *de novo* is distinguishing between isobaric amino acid residues leucine and isoleucine. Fragmentation at the peptide bond, which generates b- and y-ion series as discussed previously, cannot distinguish between the two residues.Mass spectrometry-based approaches have tentatively proposed a promising solution, by utilising high-energy CID fragmentation to generate satellite ions (Lebedev *et al.*, 2014; Zhokhov *et al.*, 2017), which are formed from cleavage within the amino acid side chain. However, these ions are often low intensity and the technique is often lacking in robustness required to make residue calls. The majority of protein identification utilises genome data as a database source, circumventing the requirement to make protein-level residue distinctions and consequently there is less demand for a protein-level solution. Previous research at the Centre for Proteome Research has employed dietary stable isotope labelling (Claydon *et al.*, 2012) to measure protein turnover in secreted proteins, and was employed as a more robust method for leucine and isoleucine distinction, despite the more laborious and expensive approach (Loxley *et al.*, 2017); this approach is discussed in further detail in chapter 3.

1.5.4 Protein quantification

Protein identification is a crucial step in most analyses of proteinaceous samples, but there is considerable importance in the abundance of identified proteins; for example, the relative ratios of MUPs are believed to be significant in the conveyance of signals that indicate identity (Roberts *et al.*, 2018). Additionally, similar abundance patterns of proteins may indicate the same biological source; urinary scent marks cannot always be distinguished from glandular scent marking, and similar expression patterns to those of proteins known to originate from certain tissues could help explain the biological origin of the proteins.

There are a multitude of methods for the quantification of proteins. Targeted mass spectrometry, commonly using a triple quadrupole mass spectrometer to isolate the peptide and its diagnostic fragment ions is utilised for its selectivity. However, quantification using isotope labelling is considered the gold standard, since the advance of instrumentation capable of distinguishing between isotope masses.

Absolute quantification of a small number of proteins can be achieved by spiking-in isotope-labelled protein or peptide standards (for example AQUA peptides (Gerber *et al.*, 2003)), and approaches such as QconCAT technology have extended this capability for

quantification of multiple proteins (Pratt *et al.*, 2006) including MUPs (Beynon *et al.*, 2014). Protein standard absolute quantification (PSAQ) is another method for absolute quantification, using an isotope-labelled protein identical to the full-length target protein, therefore behaving as closely to the analyte protein in terms of digestion, fractionation and MS analysis (Brun *et al.*, 2007). Nevertheless, these methods are suitable for analyses of proteins that have been previously identified and characterised; they are often combined with targeted mass spectrometry approaches such as multiple reaction monitoring (MRM) (Brun *et al.*, 2009), and are unsuited to global proteomic analyses.

Stable isotope labelling in amino acid culture (SILAC) is a commonly used technique for relative protein quantification in samples from cultured cell lines (Ong *et al.*, 2002), but is unsuitable for experiments on mammalian systems. Isobaric tags such as TMT can also be employed in the quantification of proteins in large, complex datasets (Thompson *et al.*, 2003). Isobaric tagging involves the attachment of mass tags to the peptides within a sample. The mass tags are identical in chemical structure and overall mass, but are cleaved within the mass spectrometer during high-energy collision. This releases a reporter ion containing substituted isotopes, such that each reporter ion can have a unique mass. It is therefore possible to label different replicates or conditions with different reporter ions and combine the sample to produce a multiplexed analysis. Newest TMT labelling kits can compare up to 16 samples in one analysis (https://www.thermofisher.com/). After labelling, samples are pooled allowing all conditions or replicates to be analysed together in the same MS run. In large datasets, off-line chromatographic fractionation is often employed after pooling the samples to reduce complexity for improved identification. During fragmentation of the precursor ion in the MS, the reporter ions are cleaved, and from their resolution in the lower *m/z* range can be used to quantify the associated peptide relative to other experimental conditions. Advance in instrumentation to perform MS3 fragmentation has also improved this quantification of reporter ions (Cao, Evans and Robinson, 2015) The multiplexed approach improves reproducibility and precision and reduces technical variability and MS time required (Li *et al.*, 2012), but can be an expensive approach.

Label-free quantification (LFQ) on the other hand, is based on the relationship between protein abundance and peptide ion intensity. Quantification can be performed on data-independent acquisition spectra (DIA), or data-dependent acquisition (DDA). DIA has an advantage in that the method overcomes the issue of dynamic range by scanning all peptides within defined *m/z* windows; however in most cases it does require pre-

acquisition of samples for generation of spectral libraries (He *et al.*, 2019). Popular methods based on DDA include intensity based absolute quantification (iBAQ) in which all identified peptide intensities are summed and divided by the number of theoretically observable peptides (Schwanhäusser *et al.*, 2011; Arike *et al.*, 2012). Other DDA label-free quantification techniques include spectral counting and TopN. Spectral counting approaches, reviewed in Bantscheff (2012), are based on the observation that the number of PSMs, peptides or total fragment ion intensity obtained for a protein is correlated with abundance (Bantscheff *et al.*, 2012). Different normalisation methods are suggested for improving the accuracy and reproducibility of lspectral counting, but compared to labelled approaches, it is one of the least accurate approaches (Choi *et al.*, 2012). In TopN, a number of confidently-identified peptides, usually the top three, are used to represent the overall protein abundance, by integrating the ion intensities of a peptide over its chromatographic elution (Bantscheff *et al.*, 2012). Intensity-based methods give better overall performance than spectral counting (Choi *et al.*, 2012), however peptides can differ in their ionisation efficiencies and it lacks accuracy in comparison to labelled approaches.

Whilst LFQ is considered the simplest quantification method, as it does not rely on chemical labelling or standards, it does come with caveats. Stable and reproducible chromatographic separation and MS acquisition is paramount, as is careful sample preparation that minimises variation. However, there is no limit to the number of samples analysed in one experiment, nor are there expensive sample preparation techniques. It is therefore a widely used technique to assess changes in protein abundance between experimental groups (Bantscheff *et al.*, 2012).

1.5.5 Glycoprotein analysis

Post-translational modifications are a crucial aspect of molecular function in proteins, and identification of modified residues can be important in determining the function, regulation or activity state of a protein. Proteomics techniques are adept at identifying common small moieties such as oxidation, phosphorylation and acetylation due to their distinctive molecular weight differences (Olsen and Mann, 2013). Fragment ions are employed to localise small PTMs within a peptide, and a number of software tools are available to confidently assess this. Glycosylation is a post-translational modification characterised by the addition of a carbohydrate molecule, usually to asparagine (*N*-linked) or to the oxygen atom in the R group of serine or threonine (*O*-linked). A number of lipocalins expressed in

scent secretions have been identified as glycoproteins, or contain an N-linked glycosylation site N-X-[S/T], suggesting that the attachment of the moiety may play a role in chemosignalling (Mechref *et al.*, 2000; Spinelli *et al.*, 2002; Watson *et al.*, 2007).

Carbohydrates can have diverse structures, forming both linear and branched chains and their size presents a similar analytical challenge as that of proteins. However, whilst the linear transmission of information from DNA to protein synthesis can provide proteomics analyses with a genome template with which to match experimental data, no such template exists to assist with glycan identification. Analysis of glycopeptides can therefore be analytically difficult, as glycan residues add additional layers of complexity. When focusing on the analysis of the protein component of a glycoprotein, attached glycans can prevent identification of the associated peptide when using standard bottom-up proteomics. Sequencing of the glycan itself also remains a challenge, although there are advances in techniques addressing complexity of the glycan component of protein samples (Rahman *et al.*, 2014). Identification of glycopeptides is accounted for in some search engines, but the additional complexity that the carbohydrate adds not only increases the search space required, but often reduces the probability of making an accurate identification (Hu *et al.*, 2016).

Removal of the attached glycan is an appropriate step forward, and a number of techniques exist. For removal of *N*-linked glycosylation, an enzymatic method using *N*-glycosidase F (PNGase F) is the most common, which cleaves the glycan from the modified asparagine and converts it to aspartic acid (Maley *et al.*, 1989). When this is incorporated into a proteolytic digestion protocol, detachment of the glycan moiety is likely improve the ionisation efficiency of both the glycan and peptide in downstream mass spectrometry analysis, for characterisation of either entity.

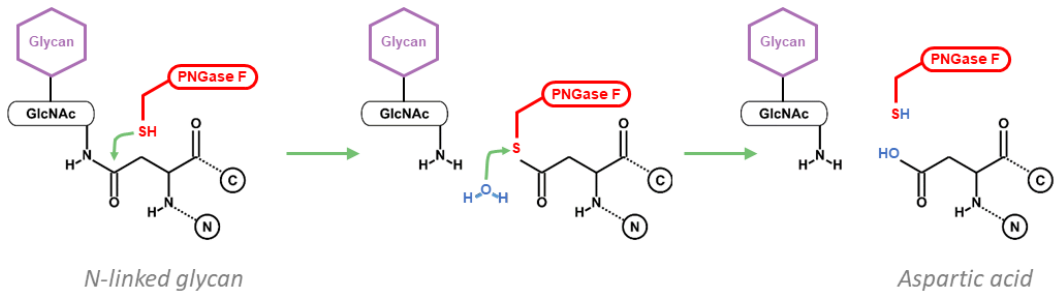


Figure 1.11 | *Action of PNGase F.*
PNGase F removes N-linked glycan moieties between the amino acid and first GlcNAc molecule, converting the asparagine residue to aspartic acid. Basic structures drawn in ChemSpider (Royal Society of Chemistry, https://www.chemspider.com/StructureSearch.aspx).

Whilst there is a range of techniques available for identification and quantification of the different components of scent marks, chemical characterisation of a scent mark is only one aspect of investigating a putative semiochemical or pheromone. Isolation, or synthetic production of the molecules is required for use in bioassays to test the elicited behaviours before a molecule is determined to have pheromonal activity. Bioassays should perform a reliable measure of the investigated behaviour, by using the semiochemical at a natural concentration. Furthermore, some semiochemicals are active as part of a blend of molecules. For example, some insects make use of pheromone blends, where the response is only elicited in the presence of all components (Wyatt, 2014), and in the case of aphrodisin, both protein and ligand as a complex is important for eliciting a behavioural response (Briand *et al.*, 2000; Briand *et al.*, 2004; Briand, Trotier and Pernollet, 2004). It is therefore important to test the components both in isolation, and also in conjunction with the other molecules comprising the scent mark.

1.6 Scope of this thesis

Evidence of olfactory communication extends across mammalian organisms, however investigation into the chemical constituents of these scents remains slow, particularly in considering the importance of non-volatile cues. The prevalence of proteins in scent-marking secretions appears to vary between species, and profiles differ from the highly polymorphic expression profiles of rats and mice to the single expression of a predominant protein such as aphrodisin and roborovskin. It has been suggested that MUP expansion is due to social organisation of these species; rats and mice have evolved to co-exist in human-populated areas at high densities, and may require a more complex system of communication that reflects this (Beynon *et al.*, 2008). However, protein expression in the scent marks of rodents with contrasting social organisations has yet to be fully addressed.

Furthermore, the expression of lipocalin sequences in scent secretion across a wider range of mammals has been suggested, but rarely investigated at the protein level. The work described in this thesis is therefore a first step in the exploration of the protein complement of scent secretions, with the goal of adding new knowledge for particular species, but also, in learning strategies for efficient acquisition of these data.

Investigation into the scent mark proteome of three species is introduced: the bank vole (*Myodes glareolus*), the field vole (*Microtus agrestis*) and the common brushtail possum (*T.vulpecula*). Both vole species were investigated in collaboration with the Mammalian Behaviour and Evolution Group, University of Liverpool, and proteomics of brushtail possum urine was performed in collaboration with Victoria University of Wellington.

Aims & Objectives

1. *To investigate expression patterns of urinary proteins across a number of mammalian systems.*

The importance of proteins in scent signalling is becoming increasingly appreciated, but there is little research into the functional protein content of scent marks beyond that of the house mouse. Initial investigation into other species indicates differential protein expression in scent marks, in terms of polymorphism and complexity, but the extent to which this has been investigated is limited. Protein synthesis is metabolically expensive, so high levels of secreted protein are taken as indicative of an important function. Therefore, an initial comparison of protein abundance and complexity within the scent marks of different species could indicate the extent of protein expression, and if further investigation into protein content may elicit discovery of novel pheromones.

2. *To fully characterise the proteins identified using mass spectrometry and other protein chemistry techniques, and furthermore assess the likelihood of their involvement in olfactory communication.*

Proteins of interest such as those with high abundance, or those that are differentially expressed between the sexes or seasons, would require further investigation to assess putative functions. Therefore, a comprehensive, biochemical assessment of each such protein was made, to understand the range of protein species expressed, and form hypotheses regarding their potential role. None of the species under investigation has an annotated genome database from which a comprehensive global proteomics search could occur, and therefore full characterisation of proteins requires isolation and potentially sequencing *de novo*. Comparison of protein sequences and structure to those proteins known to have semiochemical function can help suggest putative roles.

3. *To explore the effects of seasonality and sex on urinary protein expression, and consequently predict the role these proteins may play in communication between conspecifics.*

Many species exhibit sexual dimorphism, in both behaviour and chemical profile, in terms of scent communication. Comparison of identified proteins, both qualitatively and quantitatively, between males and females could indicate a potential semiochemical candidate. Likewise, some species exhibit differential chemical communication correlated with season or age, and a preliminary exploration of differences in scent mark protein expression could help form hypotheses regarding the potential signal conveyed.

4. *To compare patterns of protein expression in different evolutionary lineages, and consider the relationship between the behavioural biology and the urinary proteome.*

Investigation into scent mark protein expression so far indicates that murid rodents generally use both MUP-like and OBP-like proteins for chemical communication, in contrast to cricetid rodents, which tend to express OBPs like aphrodisin. However, exploration of a greater range of species is required to make a more detailed assessment. Additionally, this distinction in protein-mediated olfactory communication may extend beyond the rodent family. The brushtail possum is already reported to express a MUP-like protein from mammary gland secretions, suggesting that lipocalins are utilised in secretions across all mammalian lineages, but further insight into protein expression of a scent mark, in this case urine-mediated, may provide an initial understanding of how these proteins have evolved across mammalian lineages.

2 Experimental Strategy

2.1 Abstract

A similar, proteomics-based strategy was taken for each project. None of the species investigated had an available annotated genome, so a combination of cross-species matching and manual data interpretation were used to make identifications. Initially, overall protein content and complexity was explored, and abundant proteins displaying sexual dimorphism or seasonal changes in abundance were identified and purified. A multiple protease approach, combined with dietary isotopic labelling to discriminate between leucine and isoleucine residues, was used to sequence proteins *de novo*. Phylogenetic analyses and structural homology modelling of the novel protein sequences were employed to explore putative functionality. The sequences were also added to the databases of homologous sequences used for cross-species identification, after global proteomics of the urine and scent mark samples. A label-free quantification approach was used to identify patterns in protein expression, further investigating the effects of seasonality and sex on protein expression.

2.2 Overview

Although each project was at a different stage (for example the amino acid sequence of glareosin from bank vole urine was previously established), a similar strategy was taken for each project (Figure 2.1). Samples were usually collected by collaborating laboratories, and sent to the Centre for Proteome Research for proteomics-based analyses.

Initially, overall protein content and complexity was explored. Urine concentration was first measured based on the urinary output of creatinine, a metabolite excreted at a constant rate from skeletal muscle and exclusively expelled from the bladder, which is therefore a measure of urinary dilution (Beynon and Hurst, 2004). Protein concentration was therefore normalised as a proportion of creatinine output. Both native and SDS-PAGE, and intact mass profiling using ESI-MS, were employed to assess effects of sex and seasonality on complexity of samples.

In-gel digestion of isolated protein bands from PAGE analysis were analysedby either MALDI-ToF mass spectrometry or LC-MS/MS. Peptide mass fingerprints (PMFs) generated from MALDI-ToF were used to evaluate heterogeneity of protein bands between individuals, and LC-MS/MS data were used to identify proteins by searching databases of homologous protein sequences, with manual review.

Proteins of interest, for example those with high abundance, sexual or seasonal dimorphism, or homology to known lipocalins with scent signalling functions,were purified using off-line anion exchange chromatography, and a multiple protease approach generating overlapping peptides was used to sequence proteins *de novo*. For two species, dietary isotopic labelling was used to discriminate between isobaric leucine and isoleucine residues. Completed novel sequences were compared to homologous sequences in phylogenetic analyses and structural homology modelling, to explore putative functionality.

The novel sequences were also added to the databases of homologous sequences used for cross-species identification in a global proteomics approach of the urine and scent mark samples. A label-free quantification approach was used to identify patterns in protein expression, further investigating the effects of seasonality and sex on protein expression.

As the approaches for each project were similar, methods described below can be considered applicable to all projects. In each results chapter, methods used will reference the relevant section in this chapter, and specify any differences made for that particular project.

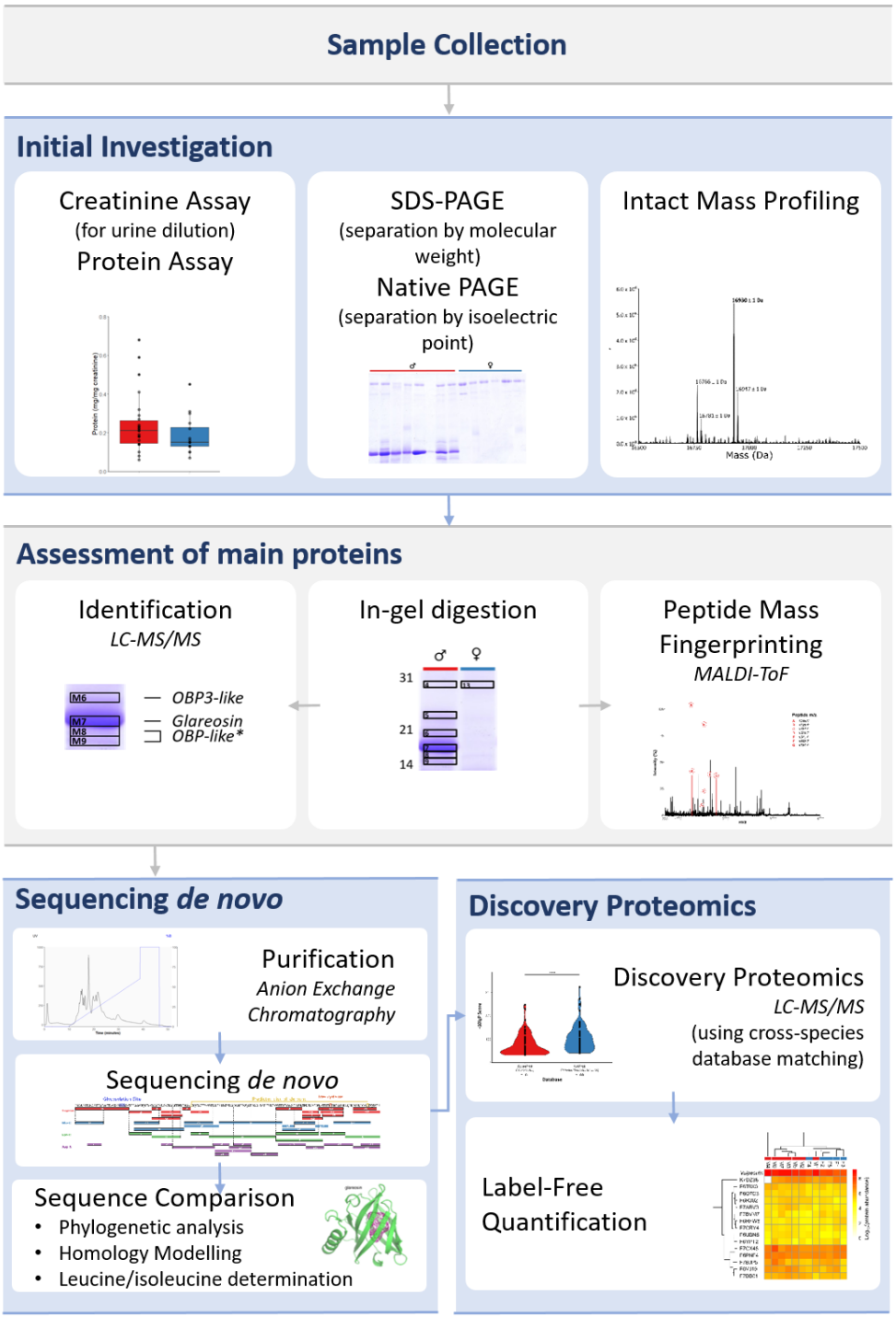


Figure 2.1 | *Experimental strategy*.
A similar strategy was used for each project. Overall protein content and complexity was assessed by protein assay, PAGE and mass profiling. The major proteins separated by PAGE were isolated and digested in-gel to be identified by LC-MS/MS, and the heterogeneity of protein bands was assessed by peptide mass fingerprinting. Identified proteins of interest, for example those with high abundance, sexual or seasonal dimorphism, or homology to known lipocalins with scent signalling functions, were purified by anion exchange chromatography and sequenced using a multiple protease approach. Leucine and isoleucine residues were determined using dietary isotopic labelling. Novel protein sequences were used for phylogenetic analysis and structural homology modelling, in addition to aiding peptide identification in discovery proteomics experiments analysing whole urine content. Label-free quantification analysed protein abundance patterns in the sequenced proteins, in addition to proteins identified by cross-species matching.

2.3 Sample Collection

Animal origin, housing methods and sample collection varied depending on species, and are therefore mainly e specified in the individual methods section for each chapter. Housing and sample collection of bank vole and field vole urine, scent mark, preputial gland and bladder urine samples followed the same protocols, as follows.

Preputial glands and bladder urine were dissected from animals culled due to a general deterioration in health caused by old age.

2.3.1 Bank vole and field vole housing conditions

All animals were housed on a reversed 16 h : 8 h "summer" light cycle in a temperature and humidity controlled environment, maintained on Lignocel® wood fibres substrate with paper wool nest material and fed on LabDiet® with food and water provided *ad libitum*. *M. glareolus* were singly housed in (43 x 11.5 x 12 cm, M3, North Kent Plastics, UK) and *M. agrestis* were singly housed in (40 x 24 x 12cm, MB1, North Kent Plastics, UK). *M. glareolus* were supplemented with mixed seeds, vegetables and berries twice a week, and *M. agrestis* were supplemented with fresh grass daily. Cardboard tubes and boxes were provided for home cage enrichment.

2.3.2 Bank vole and field vole urine collection

Voles were transferred using non-aversive handling procedures (Gouveia and Hurst, 2017) under red light conditions to clean cages (40 x 24 x 12cm) with a mesh floor to allow freely-expressed urine to pass through and collect on a plastic floor. Animals were left for up to 2 hours with regular checks for urine, which was collected using a pipette and stored in 1.5 mL plastic microcentrifuge tubes at -20°C.

2.3.3 Bank vole and field vole scent mark collection

Clean plastic Petri dishes were placed in the home enclosure for 12-24 hours. Scent marks were recovered using a cotton bud soaked in 50 µL phosphate-buffered saline (PBS) solution. Cotton buds were then placed in a 500 µL plastic microcentrifuge tube with a hole in the bottom. This was then placed in 1.5 mL plastic microcentrifuge tube and centrifuged at 8000 rpm for 2 minutes to collect the liquid. As female bank voles seldom scent mark, scent marks were recovered during urine collection whenever female scent marking occurred. Female scent marks are deposited as small sticky spots so could be distinguished from excreted urine which is deposited in larger pools. Scent marks were collected fromtwo2 captive-bred adult female *M. glareolus*.

2.3.4 Bank vole and field vole preputial gland collection

Preputial glands were dissected out of males and stored in 300 µL PBS at -20°C.

2.3.5 Bank vole and field vole bladder urine collection

Bladder urine was collected directly from the bladder during dissection using a 2.5 mL syringe. Due to the small volume retrieved, 30 µL MilliQ-grade water was added to the sample before storage at -20 °C.

2.4 Initial Assessment

2.4.1 Protein assay

Total protein concentration was measured using a Coomassie Plus protein assay kit (Pierce, Rockford, USA) using bovine serum albumin (BSA) as standard. BSA (2 mg) (Sigma-Aldrich) was reconstituted in the experiment buffer to a concentration of 2 mg/mL and was diluted accordingly to generate standards in the range of 0-30 µg/mL. To ensure samples were within the linear range of the assay, urine samples were diluted with MilliQ-grade H_2O and chromatography fractions were diluted with the appropriate buffer. Absorbance was measured at 620 nm using a microplate photometer (Thermo Fisher Scientific™ Multiskan™ FC with Thermo Scientific™ SkanIt™ Software). A linear standard curve was generated from the absorbance readings of the BSA standards, from which the protein concentration of unknowns was calculated.

2.4.2 Creatinine assay

Creatinine concentration was measured by the alkaline picrate assay kit from Sigma-Aldrich (UK). Creatinine stock solution (30 µg/mL) was diluted to provide standards in the range of 0-30 µg/mL. Samples were diluted as above. Absorbance was measured at 570 nm using a microplate photometer (Thermo Fisher Scientific™ Multiskan™ FC with Thermo Scientific™ SkanIt™ Software). A linear standard curve was generated from the absorbance readings of the BSA standards, from which the protein concentration of unknowns was calculated.

2.4.3 Polyacrylamide gel electrophoresis

PAGE analysis used a 15% Tris-chloride/Tris glycine discontinuous buffer system as standard (Laemmli, 1970). Electrophoresis was conducted in a mini-protean system (Bio-Rad, Hemel Hempstead, UK) and separated proteins were visualized with Coomassie blue. Native PAGE was carried out under non-reducing conditions. SDS-PAGE was carried out under reducing

conditions; sample buffer contained additional dithiothreitol and SDS, buffers contained SDS and samples were heated at 95 °C for 5 minutes.

2.5 Protein digestion

2.5.1 In-gel digestion

Pieces of PAGE gels containing protein bands were washed repeatedly with 25 mM NH_4HCO_3, MeCN (50:50) for 15 min at 37°C until the gel pieces were fully destained. The gel plugs were then reduced in dithiothreitol (10 mM) at 60 °C for 1 h. The dithiothreitol solution was discarded and cysteine residues were alkylated with iodoacetamide (25 µL, 55 mM) in the dark at room temperature for 45 min. The gel pieces were then washed in 25 mM NH_4HCO_3 for 15 minutes at 37 °C prior to dehydration in acetonitrile for 15 min at 37°C. Proteolytic enzymes – trypsin, endoproteinase LysC, endoproteinase GluC or endoproteinase Asp-N (each 12.5 µL, 12.5 ng/µL) were added to each of the gel pieces and incubated for 16 hours at 37°C. The digestion was stopped by the addition of formic acid (final concentration 1% v/v). Samples were centrifuged for 10 minutes at 13500 rpm prior to downstream mass spectrometry analysis.

2.5.2 In-solution digestion

Samples, diluted in 25 mM NH_4HCO_3 to 10 µg protein in a 50 µL digest were incubated with RapiGest™ SF Surfactant (0.05% w/v final concentration, Waters, Manchester, UK) at 80 °C for 10 min. The samples were then reduced with dithiothreitol (3 mM final concentration) at 60 °C for 10 min followed by alkylation with iodoacetamide (9 mM final concentration) in the dark at room temperature for 30 min. The proteases, either trypsin (0.2 µg/µL diluted in 25 mM NH_4HCO_3) (Roche, Lewes, UK), glu-C (0.2 µg/µL diluted in 25 mM NH_4HCO_3) (Roche, Lewes, UK), endoproteinase LysC (0.1 µg/µL diluted in 50mM Tris HCl, 2 mM EDTA pH 8.5), all added at a substrate:enzyme ratio of 50:1, or Asp-N (4 µg/µL diluted in 50 mM NH_4HCO_3 pH 8.0), added at a substrate:enzyme ratio of 20:1, were left to incubate for 16 hours. Following incubation, a small portion (5 µL) of the digest solution was removed and analysed by SDS-PAGE (as described in 2.4.3) to check for complete digestion. The rest of the digest solution was treated with trifluoroacetic acid (TFA) to a final concentration of 0.5% v/v and incubated at 37 °C for 45 min to precipitate the RapiGest™ SF Surfactant prior to LC-MS analysis. The samples were then centrifuged at 10,000 rpm for 15 min and the supernatant transferred to a fresh 0.5 mL plastic microcentrifuge tube.

2.6 Mass Spectrometry Techniques

2.6.1 Electrospray-mass spectrometry of intact proteins

Urine samples were diluted in acetonitrile 5% (v/v) and formic acid 0.1% (v/v) in HPLC grade water to a protein concentration of 1 pmol/µL unless specified otherwise. The samples were injected onto a C4 desalting trap (Waters MassPREP ™ Micro desalting column, 2.1 x 5 mm, 20 µm particle size, 1000 Å pore size) (Waters, Manchester, UK) that was fitted on a Waters nano ACQUITY Ultra Performance liquid chromatography® (UPLC®) system. The chromatography system was coupled to a Waters SYNAPT™ G2 QTof mass spectrometer or Waters SYNAPT^{TM} G2Si QTof mass spectrometer fitted with an electrospray source. Protein was eluted over a 10 min acetonitrile (ACN) gradient (5-90% (v/v)) at 25 µL/min. Data were collected between *m/z* 500 – 3000. The data were processed using maximum entropy deconvolution (MAX ENT 1, Mass Lynx version 4.1, Waters) at 5 Da/channel over an initial mass range of 5000-10000 Da. Once an approximate mass was determined, deconvolution of spectra was employed at 0.5 Da/channel over a smaller mass range. The mass spectrometer was calibrated externally with horse heart myoglobin (250 fmol/µL, Sigma).

Where specified, spectra were aligned using SpecAlign 2.4.1 (Oxford, UK) (Wong, Cagney and Cartwright, 2005). Peak picking was performed manually to select peaks and an average reference spectrum was generated. Baseline was subtracted with a window size of 20 Th and the spectra were cropped to focus on *m/z* ranges containing abundant peaks in multiple biological replicates of the experiment. The total ion current (TIC) was used to normalise spectra. Alignment of spectra was performed using the peak alignment by Fast Fourier Transform (PAFFT) correlation method (Wong, Durante and Cartwright, 2005). The minimum segment size was set to 1/20 the length of the spectra under analysis, with a maximum shift set to 2. The scale was set to 1 and the averaged spectrum was used as a reference.

2.6.2 MALDI-ToF Mass Spectrometry

The peptide mixtures from in-gel digestion were analysed by MALDI-TOF-MS on a Bruker ultrafleXtreme™ mass spectrometer in reflectron mode. Samples in solution were mixed with MALDI matrix solution (saturated solution of α -cyano-4-hydroxycinnamic acid in 50% (v/v) ACN / 0.2% (v/v) TFA) in a 1:1 ratio and spotted onto a target plate before being left to air dry. The laser frequency was 1000 Hz, laser energy 27% of maximum and 500 laser shots were collected per spectrum, between *m/z* 800-4000. Data were analysed in flexAnalysis v3.3 (Bruker Daltonics, Billerica, US). Peak Detection Algorithm used was Snap and a signal

to noise threshold was 6 for peak annotation. Spectra were smoothed using Savitzky Golay using a width of 0.2 Th and 1 cycle. The TopHat algorithm was used for baseline subtraction.

2.6.3 Tandem mass spectrometry

Peptides were analysed using a Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific, Hemel Hempstead, UK) coupled to a Q Exactive™ or Q Exactive™ HF Hybrid Quadrupole-Orbitrap™ mass spectrometer (Thermo Fisher Scientific, Hemel Hempsted, UK). Peptides were loaded onto a trap column (Acclaim PepMap 100, 2 cm x 75 m inner diameter, C18, 3 µm, 100 Å) at 5 µL/min with an aqueous solution containing 0.1% (v/v) TFA and 2% (v/v) acetonitrile. After 3 min, the trap column was set in-line with an analytical column (Easy-Spray ® RSLC 50cm x 75 µm inner diameter, C18, 2 µm, 100Å) (Thermo Fischer Scientific, Hemel Hempsted, UK). Peptides were eluted by using an appropriate mixture of solvents A and B. Solvent A was HPLC grade water with 0.1%(v/v) formic acid, and solvent B was HPLC grade acetonitrile 80% (v/v) with 0.1% (v/v) formic acid. Separations were performed by applying a linear gradient of 3.8% to 50% solvent B over 35 min at 300nL/min followed by a washing step (5 min at 99% solvent B) and an equilibration step (15 min at 3.8% solvent B). The mass spectrometer was operated in data dependent positive (ESI+) mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (*m/z* 300-2000, Q Exactive™; *m/z* 350-2000, Q Exactive™ HF) were acquired in the Orbitrap with 70,000 resolution (200 Th) (Q Exactive™) or 60,000 resolution (Q Exactive™ HF) after accumulation of ions to $1x10^6$ (Q Exactive™) or $3x10^5$ (Q Exactive™ HF) target value based on predictive automatic gain control (AGC) values from the previous full scan. Dynamic exclusion was set to 15 s (Q Exactive™) or 10 s (Q Exactive™ HF). The 10 (Q Exactive™) or 7 (Q Exactive™ HF) most intense multiply charged ions ($z \geq 2$) were sequentially isolated and fragmented in the octopole collision cell by higher energy collisional dissociation (HCD) with a fixed injection time of 200ms (Q Exactive™) or 100 ms (Q Exactive™ HF) and 35,000 (Q Exactive™) or 60,000 (Q Exactive™ HF) resolution with a scan range of *m/z* 200 to 2000. The mass spectrometer was calibrated using a ready to use positive ion calibration solution from the instrument manufacturer (Thermo Fisher Scientific, Hemel Hempstead, UK). The solution contains a mixture of caffeine, MRFA, Ultramark 1621, and n-butylamine in an acetonitrile:methanol:water solution containing acetic acid (1% v/v). The mass spectrometer conditions were as follows: spray voltage, 1.7 kV (Q Exactive™) or 2.5 kV (Q Exactive™ HF), no sheath or auxiliary gas flow; heated capillary temperature, 250 °C; normalised HCD collision energy 28% (Q Exactive™) or 30%

(Q Exactive™ HF). The MS/MS ion selection threshold was set to 1 x 10^4 counts (Q Exactive™) or 9.14 x 10^4 (Q Exactive™ HF) and a 3 Th (Q Exactive™) or 1.2 Th (Q Exactive™ HF) isolation width was set. Differences between parameters for each instrument is summarised in Table 2.1.

Table 2.1 | *ESI-MS/MS instrument settings.*
Differences in the established methods used for analysis of LC-MS/MS data using the Q Exactive™ and the Q Exactive™ HF mass spectrometers from Thermo Fischer Scientific.

Setting	**Q Exactive™**	**Q Exactive™ HF**
Chromatography system	Ultimate 3000 nano-LC system	
Trapping column	Acclaim PepMap 100 2 cm x 75 µm inner diameter, C18, 3 µm, 100Å	
Analytical column	Easy-Spray ® RSLC 50 cm x 75 µm inner diameter, C18, 2 µm, 100 Å	
Electrospray parameters		
Spray voltage	1.7 kV	2.5 kV
Heated capillary temperature	250 °C	250 °C
Full MS parameters		
Full scan MS range	*m/z* 300-2000	*m/z* 350-2000
Resolution	70000	60000
AGC Target	$1x10^6$	$3x10^5$
Dynamic exclusion	15 seconds	10 seconds
MS/MS parameters		
TopN for fragmentation	10 (z ≥ 2)	7 (z ≥ 2)
Injection time	100 ms	100 ms
Resolution	35000	60000
Scan range	*m/z* 200-2000	*m/z* 200-2000
MS/MS ion selection threshold	$1x10^4$	$9.14x10^4$
Isolation window	3 Th	1.2 Th
HCD collision energy	28%	30%

2.7 Data analysis

2.7.1 Database searching

PEAKS™ 8.5

Raw data were loaded into PEAKS™ 8.5 (Bioinformatics Solutions Inc., Waterloo, Canada). Precursor masses were corrected and sequences were generated *de novo* from experimental data with the following parameters: precursor error tolerance ± 10 ppm, fragment ion error tolerance ± 0.01 Da and up to five candidates per spectrum were reported. A fixed post-translational modification (PTM) of carbamidomethylated cysteine residues and variable modifications of oxidised methionine residues were specified, with three variable PTMs allowed per peptide. Additional variable modifications were added at later analyses where relevant. Database searches were performed under the same error tolerance and modification parameters. No non-specific cleavages were allowed in primary PEAKS DB searches, and the maximum number of missed cleavages was set to three. Identifications were made from SPIDER searches to allow for predicted mutations for improved identification.

Databases used are specified in the methods sections of each chapter. Each database was downloaded from UniProt (Bateman *et al.*, 2017), to include sequences from homologous species. As the unidentified peptide sequences generated *de novo* by PEAKS™ were used to either support identifications, or make new identifications, protein sequences for the proteases trypsin, glu-C, lys-C and asp-N were added to each database to identify autolysis peptides and consequently reduce time searching through unidentified spectra.

Mascot

Mascot (Matrix Science Inc, Boston, US) was used to identify proteins when using Progenesis QI for Proteomics (Waters, Manchester, UK). Databases were constructed as above from UniProt. Parameters were specified as follows; fixed modifications, carbamidomethylation of cysteine; variable modification, oxidation of methionine; peptide tolerance, 10 ppm; peptide charge 2+, 3+ or 4+; MS/MS tolerance, 0.05 Da; 2 missed cleavages allowed. A decoy search was performed and a 1% FDR filter was placed on results.

2.7.2 De novo sequencing

Purified protein was digested in-solution with one of multiple proteases as described (2.5.2) to generate sets of overlapping peptides. Peptides were analysed by LC-MS/MS as

described (2.6.3). Raw data were loaded into PEAKS™ 8.5 (Bioinformatics Solutions Inc., Waterloo, Canada) with the parameters specified in Data analysis. Abundant, unidentified peptides sequenced *de novo* by the software with an average local confidence (ALC) score of greater than 55%, and a mass tag cut-off score of 50% were manually searched and aligned using overlapping sections in addition to assistance from multiple sequence alignments (2.11.2) of homologous proteins.

2.7.3 Label-free quantification

PEAKS™8.5

Results from PEAKS DB, PEAKS PTM and SPIDER searches were used for label free quantification at 1% FDR. Mass error tolerance was set at 10 ppm, and retention time shift tolerance was set at 2 minutes. The reference sample and training samples were automatically detected by the software. Filtering parameters were set according to the data, under recommendation from the software vendor. A peptide quality score was set so that only features with a fold change of less than 8 are used to calculate peptide score, as recommended by the software vendor. The average area cut-off was set at an intensity of greater than a fold change of 8 as recommended by the software vendor; only peptides with at least one label above this intensity could be considered for quantification. A minimum peptide count of three was chosen to confidently calculate abundances, and the number of confident unique peptides was set to one. Significance was calculated using the PEAKSQ method.

Progenesis QI for Proteomics

Raw files were loaded into Progenesis QI for Proteomics. Samples were aligned to a reference sample chosen by the software. Runs with alignment scores below 80% were assisted by manually selected vectors. Features with a charge between two and four and more than two isotopes were used for quantification, and the retention time shift tolerance was set to 2 minutes. Only features with a rank of three or above were exported, so that a maximum of three spectra were used for each ion. Data was then exported and searched in Mascot according to the above criteria. Once the Mascot results were imported back into Progenesis QI for Proteomics, a protein cut-off score of 25 was set. Hi-N was selected for quantification method, using the top 3 peptides to quantify each protein. Conflicting peptides were resolved where necessary (when using a database with multiple species).

2.8 Metabolic labelling

Isobaric amino acids leucine and isoleucine were discriminated by the addition of L-leucine-5,5,5-d_3 to the animals diet over a specified feeding period (Loxley *et al.*, 2017) to an estimated incorporation rate of 50%. Urine samples were collected daily during the incorporation period as specified for each species. Tryptic in-solution digestion (2.5.2) of daily samples were performed and the digests analysed by LC-MS/MS (2.6.3) to track isotope incorporation over time. Samples collected on the day with highest incorporation were digested with each of multiple proteases in parallel to best target individual leucine or isoleucine sites, and data were analysed manually in Xcalibur (Thermo Fischer Scientific, Hemel Hempstead, UK).

Incorporation into the heavy isotope-labelled peptides was calculated using a generalized reduced gradient (GRG) nonlinear solving method. Theoretical isotope patterns for each peptide, both light and heavy, were generated in MS-Isotope (Protein Prospector v5.22.1, University of California, San Francisco, US) and each isotope peak was multiplied by a factor, F, to generate a combined spectrum for which the value of F defined the proportion that the heavy leucine spectrum contributed. For example, at a proportion of 0, the combined spectrum would be identical to the light peptide isotopic profile, whereas at 1, the combined spectrum would be identical to the heavy peptide isotope abundances. At a proportion of 0.5, the heavy and light peptide isotope peaks would be contributing 50% each. For each isotope peak, the difference between the theoretical combined spectrum and the experimental MS1 data for each peptide was calculated (E_n-T_n; E_n, experimental isotope peak intensity; T_n, theoretical isotope peak intensity). The minimum value of the sum of the residuals squared, $\sum(E-T)^2$, was then calculated using Solver (Microsoft; Frontline) by changing only the value of F, the proportion of heavy isotope incorporated.

2.9 Anion exchange chromatography

Urinary proteins were separated by anion exchange chromatography. Proteins were then loaded onto a 1 mL RESOURCE Q column (GE Life Sciences) and eluted off at 1 mL/min over a 30 column volume (CV) linear gradient using an ÄKTA purifier (GE Healthcare). Buffer conditions are specified in each chapter. Absorbance at 280 nm was monitored and fractions were collected manually, before analysis by SDS-PAGE. Fractions containing the target protein were pooled for further analysis.

2.10 Clean-up and concentration of proteins and peptides

2.10.1 Strataclean capture

Samples were made up to 1 mL with milli-Q water and 10 µL Strataclean resin (Agilent, Santa Clara, US) was added. Samples were vortex mixed at 2000 rpm for 1 minute and centrifuged at 2000 rpm for 1 minute. The supernatant was removed and 1 mL milliQ water was added, followed by vortex mixing and centrifugation as above. This wash step was repeated once more and the protein concentrated on the resin was either used downstream for SDS-PAGE analysis, by loading the protein-bead mixture, or for on-bead digestion and subsequent LC-MS/MS analysis.

2.10.2 Vivaspin® membrane ultrafiltration concentration

Protein(s) were concentrated up to 10-fold using Vivaspin® ultrafiltration cartridges (Sartorius), by centrifugation at 13000 x g at 4°C until the desired volume was reached. Columns were washed prior to use by centrifugation with an appropriate buffer for 10 minutes at 5000 x g.

2.10.3 Desalting columns

Individual urine samples were desalted using 7000 MWCO 0.5 mL Zeba™ Spin Desalting Columns (Thermo Fischer Scientific, Hemel Hempsted, UK) using the following protocol. Column storage buffer was removed by centrifugation of the column at 1500 x g for 1 minute. An appropriate wash buffer (300 µL) (for desalting prior to intact mass, HPLC-grade H_2O) was added and the centrifugation step repeated. After two more repeated wash steps, the sample was added to a maximum volume of 130 µL. If the sample volume was less than 70 µL, a stacker comprising 15 µL wash buffer was added after the sample. The column was centrifuged for 2 minutes at 1500 x g and the recovered sample was collected in a fresh plastic microcentrifuge tube.

2.11 Sequence comparison

2.11.1 BLAST searching

Whole protein sequences and shorter peptide sequences were searched using BLAST® (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Non-redundant protein sequences were searched, and organism filtering can be found in individual chapters.

Protein sequences

The blastp algorithm was employed, with the following parameters. The maximum target sequences was set to 100, the expect threshold was set to 10, the word size was set to 6 and the maximum matches in a query range was set to 0. The scoring matrix employed was BLOSUM62, and gap costs were specified to Existence: 11 Extension: 1, with a conditional compositional score matrix adjustment.

Peptide sequences

Search parameters were adjusted automatically to account for short sequence lengths, as follows; expect threshold was 200000, word size was 2, the matrix used was PAM30 and no compositional adjustment was made.

2.11.2 Multiple Sequence Alignment

Sequences selected for multiple sequence alignment were first analysed by the SignalP server (Petersen *et al.*, 2011) to identify signal peptides which were removed prior to further analyses.

MUSCLE [https://www.ebi.ac.uk/Tools/msa/muscle/] (Edgar, 2004) and Clustal Omega [https://www.ebi.ac.uk/Tools/msa/clustalo/] (Sievers *et al.*, 2011) were both employed under default parameters, chosen on the manually assessed quality of alignment. JalView (Waterhouse *et al.*, 2009) was used to visualise the alignment and alignments were coloured according to percentage identity.

2.11.3 Phylogenetic Analysis

MEGA 6.06 (Tamura *et al.*, 2013) was used for evolutionary analyses. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model. Bootstrapping analysis using 500 replicates was carried out and the tree with the highest log likelihood is shown. Branches corresponding to partitions reproduced in less than 50% boostrap replicates were collapsed. All positions containing site coverage of less than 95% coverage were eliminated.

2.11.4 Structural Homology Modelling

All RCSB Protein Data Bank (PDB) structures were searched against a novel protein sequence using BLAST® (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The 6 top-scoring sequences were aligned with the novel sequence using Clustal Omega (Sievers *et al.*, 2011) and the corresponding structures used as templates. The alignments were manually adjusted to fit the structural information given in the .pdb file, and 10 models were

generated based on each template using Modeller 9.16 (Šali and Blundell, 1993). Model quality was assessed using MolProbity (Chen *et al.*, 2010) and QMEAN score (Benkert, Silvio C. E. Tosatto and Schomburg, 2008). The highest quality model was viewed and annotated in PyMOL (PyMOL, Version 2.1.0).

2.11.5 Statistical Analysis

Linear mixed-effects models for analysis of protein and creatinine data was performed in RStudio v1.1.463 (R Core Team, 2017) using the package lme4 (Bates *et al.*, 2015) which uses a restricted maximum likelihood (REML) approach.

All other statistical analyses were performed in SPSS® Statistics version 24 (IBM® Corp., 2016).

3 Characterisation of the protein content of scent secretions in the bank vole, *Myodes glareolus*

The first part of the following chapter is constructed around Loxley *et al.* (2017). Any text shaded in yellow as demonstrated here is directly from the manuscript.

Any text shaded in yellow and boxed in a dotted line as demonstrated here is from the supplementary information of the paper.

Abstract

The urine of bank voles (*Myodes glareolus*) contains substantial quantities of a small protein that is expressed at much higher levels in males than females, and at higher levels in males in the breeding season. This protein was purified and completely sequenced at the protein level by mass spectrometry. Leucine/isoleucine ambiguity was completely resolved by metabolic labelling, monitoring the incorporation of dietary deuterated leucine into specific sites in the protein. The predicted mass of the sequenced protein was exactly consonant with the mass of the protein measured in bank vole urine samples, correcting for the formation of two disulfide bonds. The sequence of the protein revealed that it was a lipocalin related to aphrodisin and other odorant binding proteins (OBPs), but differed from all OBPs previously described. The pattern of secretion in urine used for scent marking by male bank voles, and similarity to other lipocalins used as chemical signals in rodents, suggest that this protein plays a role in male sexual and/or competitive communication. We propose the name glareosin for this novel protein to reflect the origin of the protein and to emphasise the distinction from known OBPs.

Contributions

This project was a continuation of research by Dr Michael Turton and Dr Jenny Unsworth during their time at the Centre for Proteome Research, University of Liverpool. A paper was published in 2017 as a culmination of their work, in addition to the author's contribution. Contributions are as follows:

A seasonally-expressed, sex-specific protein, glareosin, was initially identified in an initial cohort of bank vole urine samples by Dr Michael Turton. Dr Turton performed initial sequencing (Turton, 2007), using a combination of Edman degradation and tandem mass spectrometry assisted by isotopic labelling with ^{18}O. Confirmation of the sequence was performed by Dr Jenny Unsworth by digesting the purified protein with multiple enzymes and employing higher-resolution tandem mass spectrometry (Unsworth, 2014). Phylogenetic analysis and homology modelling of the completed sequence of glareosin was performed by Dr Daniel Rigden. Dr Deborah Simpson performed intact mass for protein profiling of initial samples. Alexandra Jebb collected a second cohort of urine samples, the analysis of which is described in the first part of this chapter until Section 3.4.2, and contributed protein & creatinine data to the published paper Loxley *et al*. 2017.

The author confirmed the continuity of glareosin expression in the second cohort of samples, after several generations of bank voles, and completed the glareosin sequence using dietary isotopic labelling followed by urine collection and high-resolution mass spectrometric analysis to discriminate between leucine and isoleucine residues.

All authors contributed to the manuscript and gave final approval for publication. The paper has been integrated into thesis format in accordance with the rest of the thesis. The published pdf form is attached in supplementary material.

As detailed in the second section of this chapter (Section 3.4.6 onwards), additional experiments using proteomics techniques were performed by the author to investigate the urinary proteome of the bank vole beyond that of glareosin, investigating heterogeneity and sexual dimorphism via label-free quantification at the peptide-detectable level. Initial investigations of the protein output of scent marks, bladder urine and preputial glands were also undertaken using a combination of SDS-PAGE and LC-MS/MS.

Holly Coombes collected all scent marks, bladder urine and preputial gland washes for the analyses in the second part of this chapter, from Section 3.4.6 onwards, and performed SDS-PAGE analysis of these samples where specified.

Published Paper:

Glareosin: a novel sexually dimorphic urinary lipocalin in the bank vole, *Myodes glareolus*

Grace M Loxley, Jennifer Unsworth[1], Michael J. Turton[1], Alexandra Jebb[2], Kathryn S Lilley[3,#], Deborah M Simpson[1], Daniel J. Rigden, Jane L. Hurst[2] & Robert J. Beynon[1]*

[1]Centre for Proteome Research, Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB

[2]Mammalian Behaviour and Evolution Group, Institute of Integrative Biology, University of Liverpool, Leahurst Campus, Neston CH64 7TE

[3] Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB

[4] Department of Biochemistry, University of Leicester, Leicester LE1 7RH

current address, Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA

Ethical Statement

All procedures involved in this study were non-invasive. Animal trapping, use and care were in accordance with EU directive 2010/63/EU, UK Home Office code of practice for the housing and care of animals bred, supplied or used for scientific purposes, and UK research funder's guidelines on responsibility in the use of animals in bioscience research. The University of Liverpool Animal Welfare Committee approved the work, but no specific licenses were required.

Acknowledgments

We are grateful to Dr Richard Humphries, Amanda Davidson and the animal care staff at the Mammalian Behaviour and Evolution Laboratory, and Dr Philip Brownridge at the Centre for Proteome Research for excellent instrument care. J Unsworth is grateful to the Biotechnology and Biological Sciences Research Council (BBSRC) for provision of a PhD studentship. This work was funded in part by BBSRC (BB/J002631/1 and BB/M012557/1).

3.1 Introduction

Murine rodents (Old World rats and mice, sub-family Murinae) express a set of proteins known as major urinary proteins (MUPs), which can be highly polymorphic, whereas hamsters and voles (sub-families Cricetinae and Arvicolinae) seem to express chemosignalling lipocalins more typical of the odorant binding proteins (OBP) family.

Bank voles live in small, mixed sex groups during the winter that break up in the breeding season. While breeding females inhabit non-overlapping home ranges close to the over-wintering site, males have larger overlapping home ranges within hierarchical groups that overlap several females (Mazurkiewicz, 1981; Bujalska, 1990). Males deposit urine around their territories in numerous small scent marks, using long brush-like hairs on the prepuce tip to streak out their scent, in volumes independent of the total bladder volume (Johnson, 1975; Christiansen, 1980), contrasting with the excretion of urine in pools by females (Johnson, 1975; Verplancke, Le Boulengé and Diederich, 2011). Scent marking rates are particularly high in new environments, while dominant males mark subordinate male burrow and nest areas continually. Females prefer males that scent mark more frequently (Kruczek, 1997).

Relatively little is yet known about the expression of chemosignalling proteins in the vole family, but sexual dimorphism in urinary protein expression has been observed in the bank vole, *Myodes glareolus,* where protein levels are much higher in males (Kruczek and Marchlewskakoj, 1985). Three male-specific odorant binding proteins (OBPs) have been identified in male bank vole urine that might play a role in chemical signalling (Stopková *et al.*, 2010a).

To understand the expression and potential role of urinary proteins in bank vole communication further, we examined the expression of urinary proteins in wild-caught and captive-bred voles in the breeding and non-breeding season. Here, we characterise a new urinary protein in *M. glareolus*, distinct from those previously identified, that is expressed at high level only by males and only in the breeding season. The complete protein sequence was obtained primarily using in-solution protease digestion followed by tandem mass spectrometry, distinguishing between the otherwise isobaric amino acids leucine and isoleucine using metabolic labelling. Homology modelling and structural analysis reveal strong similarity to known OBPs, but this protein is distinct from those previously described in bank voles or in other species and is the most abundant urinary protein expressed by male bank voles. Given the potentially important investment by male bank voles in this

particular urinary protein during the breeding season, we propose the name glareosin to distinguish this from other OBPs.

3.2 Aims & Objectives

The importance of proteins in the scent cues from mice and rats is well defined, however the extent to which this protein mechanism translates across other mammalian systems is far less understood, even in closely related rodent species. Previous work from the collaboration of the Mammalian Behaviour & Evolution group and the Centre for Proteome Research at the University of Liverpool identified and sequenced, bar leucine and isoleucine discrimination, a seasonally expressed, sexually dimorphic protein expressed in the urine of the male bank vole, *M.glareolus.* To investigate this protein further, discrimination of leucine and isoleucine residues was required for complete sequencing. This protein, named glareosin, which dominated protein urinary profiles, differed from homologous bank vole urinary OBPs previously described in the literature, but the presence or absence of these OBP sequences had not yet been confirmed in any samples analysed at the Centre for Proteome Research, University of Liverpool. A global proteomics approach was taken to identify peptide-level based evidence of these additional sequences. Furthermore, bank vole urine marking behaviour differentiates from scent marking behaviour, but no work has yet assessed the protein component of bank vole scent marks, and investigation is required to gain insight into complexity and heterogeneity of proteins in scent marks in comparison to urine. The added work described in this chapter addresses the following objectives.

1. To complete the sequence of the previously identified predominant protein in bank vole urine, glareosin.
2. To investigate the effects of seasonality and sex on the presence and abundance of lipocalins in bank vole urine.
3. To begin an investigation into the protein content of bank vole scent marks, and to explore sex-specific differences.
4. To compare the protein expression in bank vole urine to protein semiochemistry in other species, in which role and function of the relevant proteins have already been assessed.

3.3 Methods

As a similar approach was taken for each project, all protocols are as found in Chapter 2: Experimental Strategy, unless otherwise specified below.

3.3.1 Sample collection

Housing conditions and sample collection protocols can be found in Chapter 2: Experimental Strategy, section 2.2.

Urine samples were collected from both wild caught and captive bred *M. glareolus* voles derived from two different geographic areas of the United Kingdom (Wirral peninsula in Merseyside, approximately N 53.288°, E -3.028° and Kielder Forest in Northumberland, approximately N 55.208°, E -2.528°). This was to allow comparison of samples from different populations of bank voles for a more representative analysis.

Scent marks were collected from captive bred male and female *M.glareolus* subjects, from a colony derived from wild ancestors captured from four different populations in the northwest of England, UK (Wirral peninsula in Merseyside), regularly supplemented with wild individuals to maintain a healthy genetic population.

3.3.2 Protein output

As described in Chapter 2: Experimental Strategy, section 2.3.1.

3.3.3 Polyacrylamide gel electrophoresis

As described in Chapter 2: Experimental Strategy, section 2.3.2.

3.3.4 In-gel digestion

As described in Chapter 2: Experimental Strategy, section 2.4.1.

3.3.5 In-solution proteolysis

As described in Chapter 2: Experimental Strategy, section 2.4.2.

3.3.6 Edman degradation

SDS PAGE gels before staining were electroblotted to polyvinylidene difluoride (PVDF) membranes for N-terminal sequencing using an Applied Biosystems 476A gas phase sequencer (Applied Biosystems). After electroblotting, the PVDF was stained with Coomassie blue to visualize protein bands prior to excision and Edman degradation.

3.3.7 MALDI-ToF mass spectrometry

Analysis of peptides from in-gel digests was undertaken using a M@LDI-TOF reflectron mass spectrometer (Waters, Manchester, UK) in positive ion mode. The instrument was calibrated using a peptide mix (des-Arg-Bradykinin (2.4 pmol/µL 1,904.47 Da), neurotensin (2.4 pmol/µL, 1672.92 Da), adenocorticotropic hormone clip 18-39 (2.6 pmol/µL, 2465.20 Da) and oxidised insulin β-chain (30 pmol/µL, 3495.9 Da). Samples were prepared for analysis by co-crystalisation with an equal volume of matrix solution (saturated α-cyano-4-hydroxy cinnamic acid in a 1:1:1:1 (v/v) solution of acetonitrile:H_2O:methanol:0.1 % (v/v) TFA). A 1µL aliquot of the peptide/matrix mixture from each sample was deposited onto a 96 well MALDI target and allowed to dry at room temperature. Spectra were acquired between *m/z* 1000 and 4000 with the laser energy optimized to give the best signal to noise ratio for each sample. The laser firing rate was 5 Hz and 10 spectra (collected over 2 s) were combined. The final mass spectrum was a combination of 10-15 such combined data sets, representing 100-150 individual laser shots. Each spectrum was then internally recalibrated using the trypsin autolysis peak at 2163.057 Da or by the addition of an internal calibrant, neurotensin (1672.918 Da), (120 fmol/µL). All aspects of data acquisition, processing and machine management were controlled through the MassLynx software suite (version 4.0).

3.3.8 Electrospray ionization mass spectrometry of intact proteins

All ESI-MS was undertaken on a Q-ToF Micro mass spectrometer (Waters, Manchester, U.K.) in positive ion mode. The instrument was calibrated with the product ions of a 500 fmol/µL (Glu1)fibrinopeptide B (GluFib) solution in 50% (v/v) acetonitrile/0.1% (v/v) formic acid, infused from a syringe pump at 0.5 µL/min through a PicoTip emitter (New Objective, Massachusetts, USA). For intact mass analysis the instrument was operated in ToF only mode. Samples were desalted on-line with a C4 reverse phase trap and subsequently introduced into the mass spectrometer in a solution of 90% (v/v) acetonitrile/0/1% (v/v) formic acid. Raw data were gathered between *m/z* 700 and 1400 at a scan/interscan time of 2.4 s/0.1 s. These raw data were subsequently de-convoluted using the MaxEnt 1 module contained within the MassLynx 4.0 software.

3.3.9 Tandem mass spectrometry

Initial MS/MS analysis of proteolytic peptides was undertaken on a Q-ToF Micro mass spectrometer (Waters, Manchester, U.K.) in positive ion mode, calibrated as described in section 3.3.6.

For MS/MS analysis of proteolytic peptides, precursor spectra were acquired between *m/z* 400 and 1500 at a scan/interscan time of 1.0 s /0.1 s. Product ion spectra were acquired between *m/z* 100 and 2000 at the same scan/interscan speed. Raw product ion spectra were deconvoluted using the MaxEnt 3 algorithm in the MassLynx software, with the charge state of the parent peptide determined from the isotope envelope in the precursor ion spectrum. Interpretation of product ion spectra and the determination of peptide sequences were facilitated by the PepSeq module within MassLynx 4.0. As an additional aid in the interpretation of tandem mass spectra, peptides were isotopically labelled with ^{18}O by performing proteolytic digestion in a 1:1 mix of light ($H_2[^{16}O]$) and heavy ($H_2[^{18}O]$) water. Incorporation of a 1:1 mixture of $[^{16}O]$ and $[^{18}O]$ atoms into the newly formed C-termini of peptides prior to tandem mass spectrometry, allowed y-ions to be identified as a sequence of doublets of approximately equal intensity, separated by 2 Da.

To confirm the sequence, we repeated the digestions and analysed the samples on a high-resolution instrument with high mass accuracy and resolution for precursor and product ions. For this stage, samples were analysed using a Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific, Hemel Hempstead, UK) coupled to a Q Exactive™ mass spectrometer (Thermo Fisher Scientific, Hemel Hempstead, UK).

For high-resolution Q Exactive™ mass spectrometry methods, please see Chapter 2: Experimental Strategy, section 2.5.3.

3.3.10 Database searching

As described in Chapter 2: Experimental Strategy, section 2.6.1.

3.3.11 Label-free quantification

As described in Chapter 2: Experimental Strategy, section 2.6.3.

3.3.12 Metabolic labelling

To discriminate between isobaric leucine and isoleucine residues, we fed bank voles a diet containing stable isotope labelled leucine. Low-sugar rodent diet (100 g, 1.16% w/w leucine) was suspended in milliQ-grade H_2O. $[^2H_3]$leucine (5,5,5-d3-leucine) was added (1.16 g) to a level equivalent to that present in the diet in unlabelled form and the diet and label were blended thoroughly in a food processor. The diet was then dehydrated for 48 hours in a commercial food drier. Cage deposited urine samples were collected from four voles (day 0) before they were transferred to a new cage with the $[^2H_3]$leucine diet provided *ad libitum*. Cage deposited urine samples were then collected daily for four days. Unlabelled

diet was resumed after the final urine samples were collected. Urine samples were then stored at -20°C until analysis. Urinary proteins were reduced, alkylated and digested with trypsin in solution, followed by LC-MS/MS analysis on a QExactive-HF (Thermo ScientificTM) as described above. Leucine and isoleucine residues were then manually assigned from the raw data, and confirmed with MASCOT and PEAKS searches under the same search conditions as below with triple labelling with deuterium as an additional variable modification, against the derived sequence of glareosin.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (32) partner repository with the dataset identifier PXD006645 and 10.6019/PXD006645

3.3.13 Anion Exchange Chromatography

Proteins were purified by anion exchange chromatography using a UNO-Q anion exchange column (Bio-Rad, Hemel Hempstead, UK) equilibrated in 10 mM HEPES pH 8.0 and were eluted with a 0 - 0.5 M NaCl gradient at 1 mL/min. The absorbance of the eluate was measured at 280 nm and 0.5 mL fractions were collected. The fractions were subsequently analysed by SDS PAGE and electrospray ionisation mass spectrometry (ESI-MS).

3.3.14 Protein sequence analysis:

The final amino acid sequence was used in a BLAST search (Altschul *et al.*, 1990) using default parameters for protein matches against *Rodentia*. The 138 matches were reduced and processed as follows. First, incomplete sequences, sequences substantially larger than the core lipocalin size of approx. 160 amino acids or those that only matched across part of the sequence were eliminated. Some sequence entries were exact duplicates and were reduced to single entries. Finally, because we wished to compare the glareosin secreted protein sequence, signal peptides were removed, either guided by the feature entry in the database entry or through the SignalP 4.1 server (Petersen *et al.*, 2011) (http://www.cbs.dtu.dk/services/SignalP/). The reduced sequence set was aligned with MAFFT using the high accuracy linsi algorithm (Katoh and Standley, 2013) with Jalview (Waterhouse *et al.*, 2009) used to display and manipulate sequence alignments.

3.3.15 Phylogenetic analysis:

The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones, Taylor and Thornton, 1992). Bootstrapping analysis (Felsenstein, 1985) using 500 replicates was carried out. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. All positions

containing gaps and missing data were eliminated leaving a total of 112 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Tamura *et al.*, 2013).

3.3.16 Homology modelling:

The structure of mature glareosin, without its signal peptide, was modelled using the Rosetta_CM protocol (Song *et al.*, 2013). Ten models were produced for each combination of templates and alignments. Templates were identified from a non-redundant library of PDB structures using the HHpred server (Söding, Biegert and Lupas, 2005) and modelling was done with one, five or ten templates assessing the results quantitatively with Rosetta's own energy function and with the Prosa II (Sippl, 1993), DOPE (Shen and Sali, 2006) and QMEAN (Benkert, Silvio C E Tosatto and Schomburg, 2008) protein structure quality metrics. Stereochemistry was assessed with PROCHECK (Laskowski *et al.*, 1993). Structures were superimposed using GESAMT (Krissinel, 2012). Cavities were detected and measured using the GHECOM (Kawabata, 2010) and Profunc (Laskowski, Watson and Thornton, 2005) servers. PyMOL (https://www.pymol.org/) was used to visualise and manipulate structures and to produce structure figures.

3.4 Results

3.4.1 Preliminary work

Previous assessment

Assessment of seasonal and sex variation confirmed that urinary protein output was substantially higher in males, but only during the breeding season (interaction between season and sex, $F_{1,21} = 5.19$, $P = 0.033$; Figure 3.1a). Male average protein output increased over three-fold, from 3.5 ± 0.5 mg protein / mg creatinine during the non-breeding season (uncorrected urinary protein concentration 0.36 ± 0.07 mg/ml) up to 11.2 ± 1.2 mg protein / mg creatinine in the breeding season (uncorrected urinary protein concentration 1.76 ± 0.27 mg/ml). As urinary creatinine levels were not influenced by season or sex, these differences in urinary protein output were entirely due to differences in the concentration of protein excreted in urine. A preliminary assessment of protein complexity in these samples by 1D SDS-PAGE revealed an intense band between 14 and 21 kDa that was evident only in male samples and only during the breeding season (Figure 3.1b,c). We also assessed protein output in urine samples from bank voles bred in captivity and kept under breeding season lighting conditions but without sexual experience. This confirmed a highly significant sex difference in urine protein output ($F_{1,28} = 79.6$, $P < 0.0001$), with levels comparable to those seen in wild-caught voles during the breeding season (Figure 3.1a). Thus, elevated protein output in males was not dependent on sexual experience. This elevated protein output was evident in male bladder urine, sampled when older voles were culled (effect of sex, $F_{1,10} = 6.8$, $P = 0.026$). SDS-PAGE confirmed that the same intense band between 14 – 21 kDa was present in male but not in female samples, in both naturally deposited and bladder urine (data not shown).

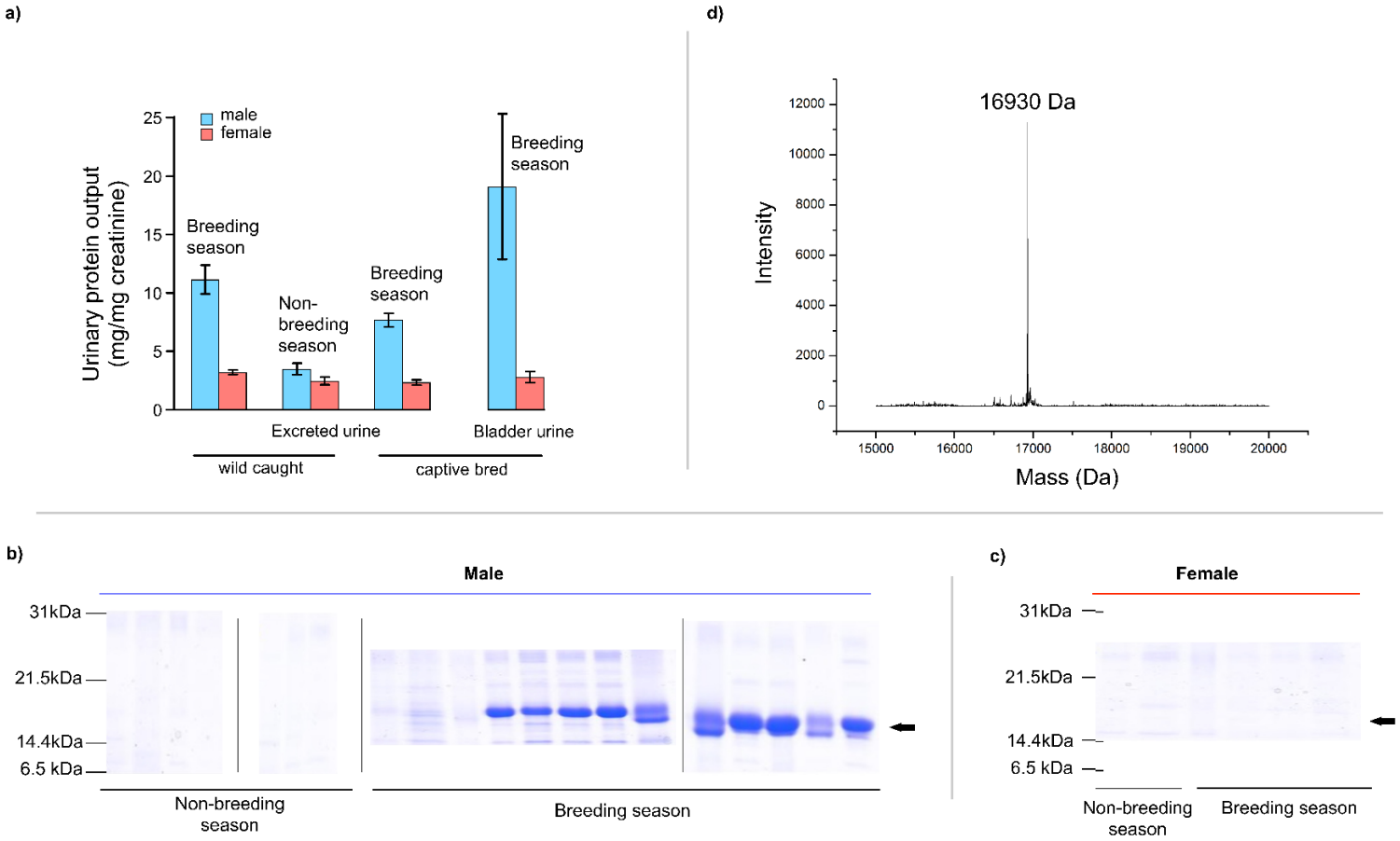


Figure 3.1 | *Analysis of bank vole urinary protein output.*

For urine samples from adult male or female bank voles (see text), protein and creatinine concentrations were determined and expressed as mg protein:mg creatinine to correct for urine dilution (a). The protein complement was analysed by SDS-PAGE (b, male and c, female; vertical bars separate discrete gels) and by electrospray ionization mass spectrometry (male, d).

Intact mass analysis has also been used to assess the heterogeneity of urinary proteins in both captive bred (Evershed et al. 1993; Robertson et al. 1996; Mudge et al. 2008; Beynon et al. 2015) and wild caught mice (Robertson *et al.*, 1997; Beynon *et al.*, 2002; Sheehan *et al.*, 2016; Hurst *et al.*, 2017), identifying small mass changes caused by discrete amino acid substitutions in the protein sequence. The intact mass profiles of the *M. glareolus* urinary proteins was analysed by ESI-MS (Figure 3.1d). A single predominant intact mass was measured in all samples at 16930 ± 1 Da and there was no evidence of inter or intra-individual heterogeneity in the mass profile. The protein identified at a mass of 16930 Da in all *M. glareolus* urine samples was purified by anion exchange chromatography. This ion exchange protein was recovered and used for primary sequence analysis, as genomic or transcriptomic data were lacking. The measured intact mass differed from the predicted masses of the three urinary OBP proteins reported in bank voles by Stopková and colleagues (Stopková *et al.*, 2010b) which, allowing for the formation of two putative disulfide bonds, together with loss of signal peptide predicted by signal (Petersen *et al.*, 2011) were OBP1 (D3VW62_MYOGA): 16643 Da, OBP2 (D3VW64_MYOGA): 16837 Da and OBP3 (D3VW62_MYOGA): 16749 Da, consistent with this being a novel protein.

After 1-D SDS PAGE and blotting to PVDF membrane, the 16930 Da protein was partially sequenced by gas phase Edman degradation. Although less commonly used today, Edman degradation permits precise positioning of the true N-terminal sequence of the protein. The recovered sequence, HSEIDEKWVTVAIAADNVNK used in searching (BlastP) (Altschul *et al.*, 1990) with standard search parameters, aligned most strongly to the N-terminal sequence of a prairie vole (*Microtus ochrogaster*) aphrodisin-like protein 1 (best match: XP_005372052; 70% identity) and a bank vole (*M. glareolus*) OBP1 (best match, D3VW62_MYOGA; 65% identity) as well as other members of the lipocalin family. This match pointed to the potential role of this urinary protein as a semiochemical lipocalin. Although the N-terminal sequence overlapped with the first structurally conserved GlyXxxTrp region of the lipocalin family (GXW, (Flower, 1996)), the highly-conserved glycine residue of the motif was absent. However, Glu (E) and Gly (G) elute in close proximity in Edman degradation, raising the possibility of a miss-call at this position.

Additional studies

A single predominant protein identified in male bank vole urine during the breeding season had been previously purified and sequenced. The next step was to confirm the expression pattern and sequence of glareosin were the same in the second cohort of samples, collected after several generations of captive bank voles, so that leucine and isoleucine residues could be resolved. To ensure protein output was consistent over the time passed between projects, bank vole urine samples (n=61) were collected from 49 mature bank voles, 21 male and 28 females, held captive in solitary conditions under a light cycle simulating breeding season conditions and analysed (Figure 3.2).

Sex affected urine dilution ($\chi^2(1)=4.6921$, $p=0.0303$); females had a lower creatinine concentration in relation to males by a value of 99.77 ± 45.78 (mean ± SE) µg/mL (Figure 3.2A). Sex does statistically affect protein output ($\chi^2(1)=23.397$, $p<0.01$), and females have a lower protein concentration in relation to males by a value of 658.12 ± 122.34 (mean ± SE) µg/mL (Figure 3.2B). Protein was recorded relative to creatinine output to correct for urine dilution, and differed significantly between sexes ($\chi^2(1)=19.031$, $p<0.01$), and females have a lower dilution-corrected protein ratio by a value of 2.56 ± 0.5440 (mean ± SE) (Figure 3.2C). The conclusions from Loxley et al. (2017) can therefore be reviewed, in reference to the effect of sex on urinary dilution.

SDS-PAGE analysis confirmed a sexually dimorphic, seasonally specific protein band resolving to approximately 18 kDa (Figure 3.2D), and native PAGE revealed a predominant protein band with several weaker bands resolving either side (Figure 3.2E).

Male urine samples from captive bank voles (n = 10) were normalised to protein (1 pmol) and analysed by ESI-MS. Spectra were analysed in MassLynx (Waters) and after baseline subtraction and smoothing using the Gaussian method, were initially deconvoluted to a wide range (5 – 100 kDa) before focussing on the region of the predominant protein (5 – 30 kDa). Spectra were aligned to a generated average spectrum in Spec Align using the PAFFT method and normalised to TIC. Spectra were overlaid. All samples were similar, with glareosin (16930 ± 1 Da) the predominant band in most samples, with the additional peak at 16947 ± 1 Da likely to be oxidation (for individual spectra see supplementary figures).

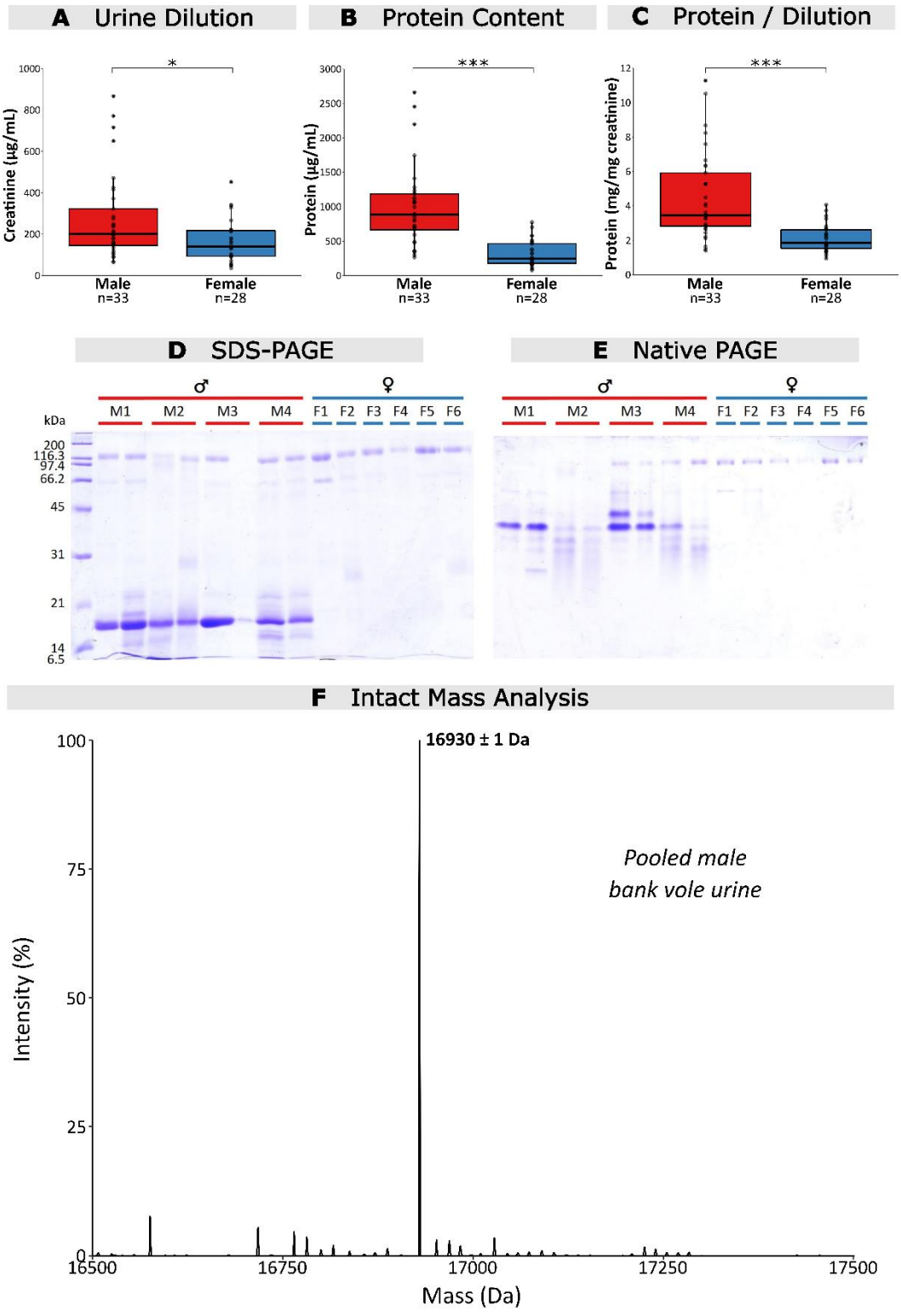


Figure 3.2 | ***Exploring consistency in bank vole urinary protein output.***
The protein content and creatinine content of urine samples from 21 mature male (n = 33) and 28 mature female (n = 28) bank voles collected under breeding season conditions were measured (A-C). Statistical analysis was performed in R using mixed effect models to correct for repeated samples from the same animal. * = $p < 0.05$, *** = $p < 0.001$. Samples normalised to 1 µg creatinine were analysed by SDS-PAGE (D) and native PAGE (E) to assess protein content further. Intact protein (1 pmol) in pooled male urine (F) and individual samples (supplementary material) was analysed by ESI-MS.

To assess if bank vole urinary content is consistent in both captive and wild bank voles, urine samples from trapped wild bank voles were collected during the winter (male samples n = 8; female samples n = 2) and during the breeding season, where bank voles were categorised as immature (males, n = 8; females, n = 2) or mature (males, n = 5; females, n = 2). Protein output and urine dilution were measured (Figure 3.3A), and although statistical analysis was omitted due to low sample numbers, protein output relative to dilution in breeding season males was far higher than female and winter male samples. SDS-PAGE of urine was normalised to 0.6 µg creatinine (Figure 3.3B) and again displayed a strong protein band at approximately 18 kDa. Ten samples obtained in the breeding season from five immature and five mature male bank voles were analysed by ESI-MS for protein profiling (Figure 3.3C). The predominant protein had a mass consistent with glareosin, allowing for instrument error, in all but one sample, and only three other peaks were observed at a lower intensity; 16766, 16781 (likely to be methylation or oxidation of 16766 Da) and 16947 ± 1 Da, likely to be the oxidised form of glareosin. This is consistent with the results observed in captive voles (see Figure 3.2) and shows the generational consistency of the bank vole urinary protein profile.

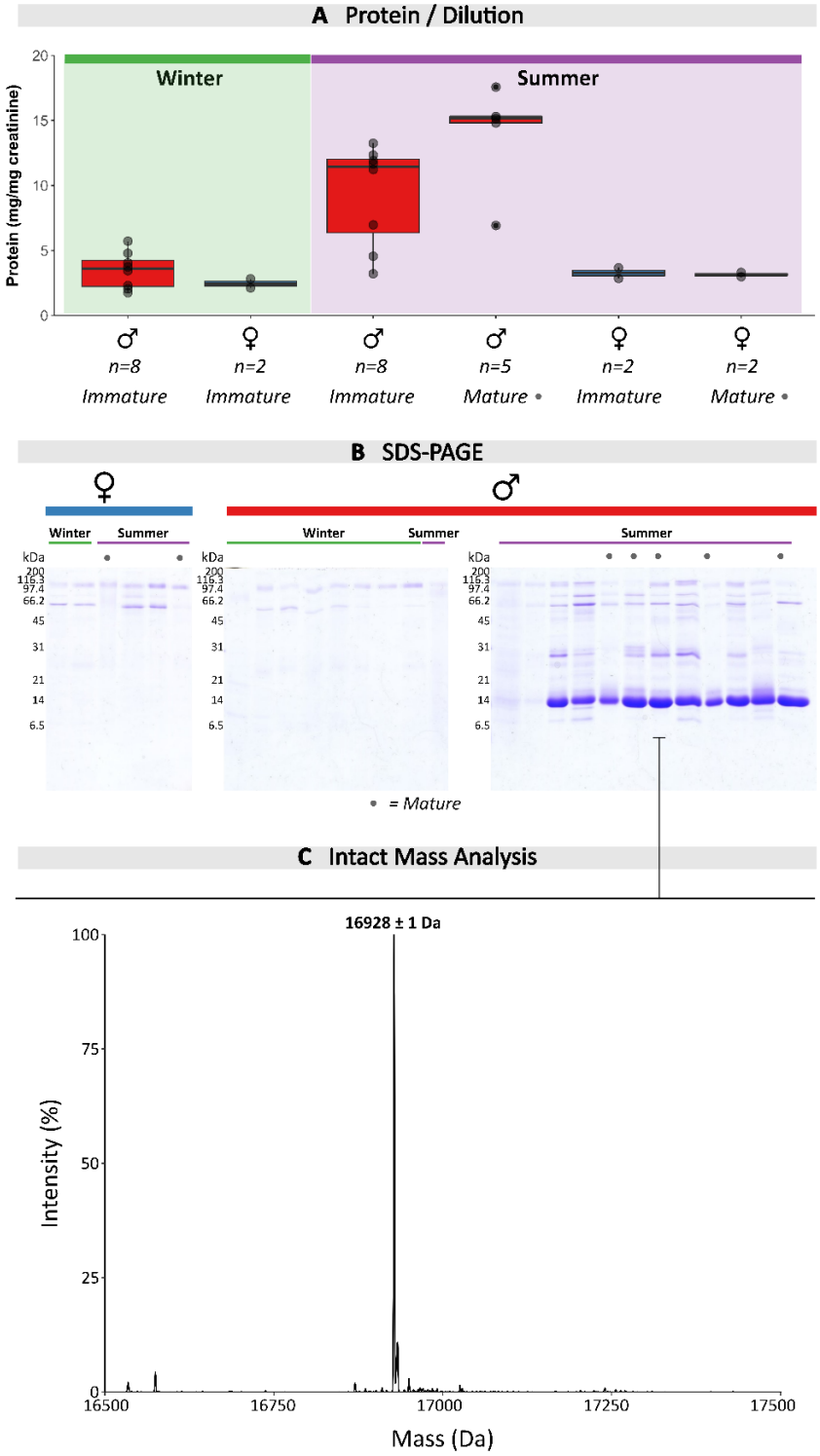


Figure 3.3 | *Exploring continuity in wild bank vole urinary protein output.*
The protein and creatinine content of urine samples from wild-caught male and female bank voles, collected in and out of the breeding season were measured (A). The weight of the bank voles was recorded and voles were assigned as mature or immature. Statistical analysis were not performed due to low sampling numbers. Samples normalised to 0.6 µg creatinine were analysed by SDS-PAGE (B) to assess protein content further. Five mature and five immature breeding season male samples (1 pmol protein) were analysed by ESI-MS and the resulting protein profiles were aligned in SpecAlign and normalised to TIC. An example of a deconvoluted spectrum from analysis of the urine from a mature male (of which the corresponding gel lane is indicated) is displayed in C.

As bank vole urinary protein profiles were consistent over time, a combined dataset of protein and creatinine output was produced from current and previous samples, to include data from outside the breeding season, in addition to bladder urine samples.

3.4.2 *Sequencing* de novo

To gain further information about the 16930 Da protein, a peptide sequencing strategy based on mass spectrometry was adopted. Two approaches were taken. First, Q-TOF tandem mass spectrometry of peptides obtained by direct infusion of proteolytic digests of the purified protein and secondly, LC-MS/MS of the peptide mixture on a second instrument that generated product ions at high mass accuracy and resolution. The sequencing strategy was based on digestion with three different endopeptidases (trypsin, endopeptidase LysC and endopeptidase GluC) to generate overlapping peptides that would cover as much of the primary sequence of the mature protein as possible, although unable to discriminate between the isobaric Leu/Ile pair, signified here by the residue "J". In some instances, interpretation of the fragment ion mass spectra was assisted by labelling peptides using a 1:1 ratio of $H_2{}^{16}O$:$H_2{}^{18}O$ in the digestion reaction. Only the y series of ions, derived from the C-terminus of each peptide, are isotopically labelled in this reaction, and the doublets thus facilitated discrimination of the b and y ion series. Following interpretation of the amino acid sequence from the fragmented peptide, the theoretical *m/z* value of the $[M+H]^+$peptide was calculated and reconciled with the ions observed by MALDI-TOF. The complete sequence strategy is presented in Figure 3.4 and the relevant peptide mass spectra are presented in supplementary materials.

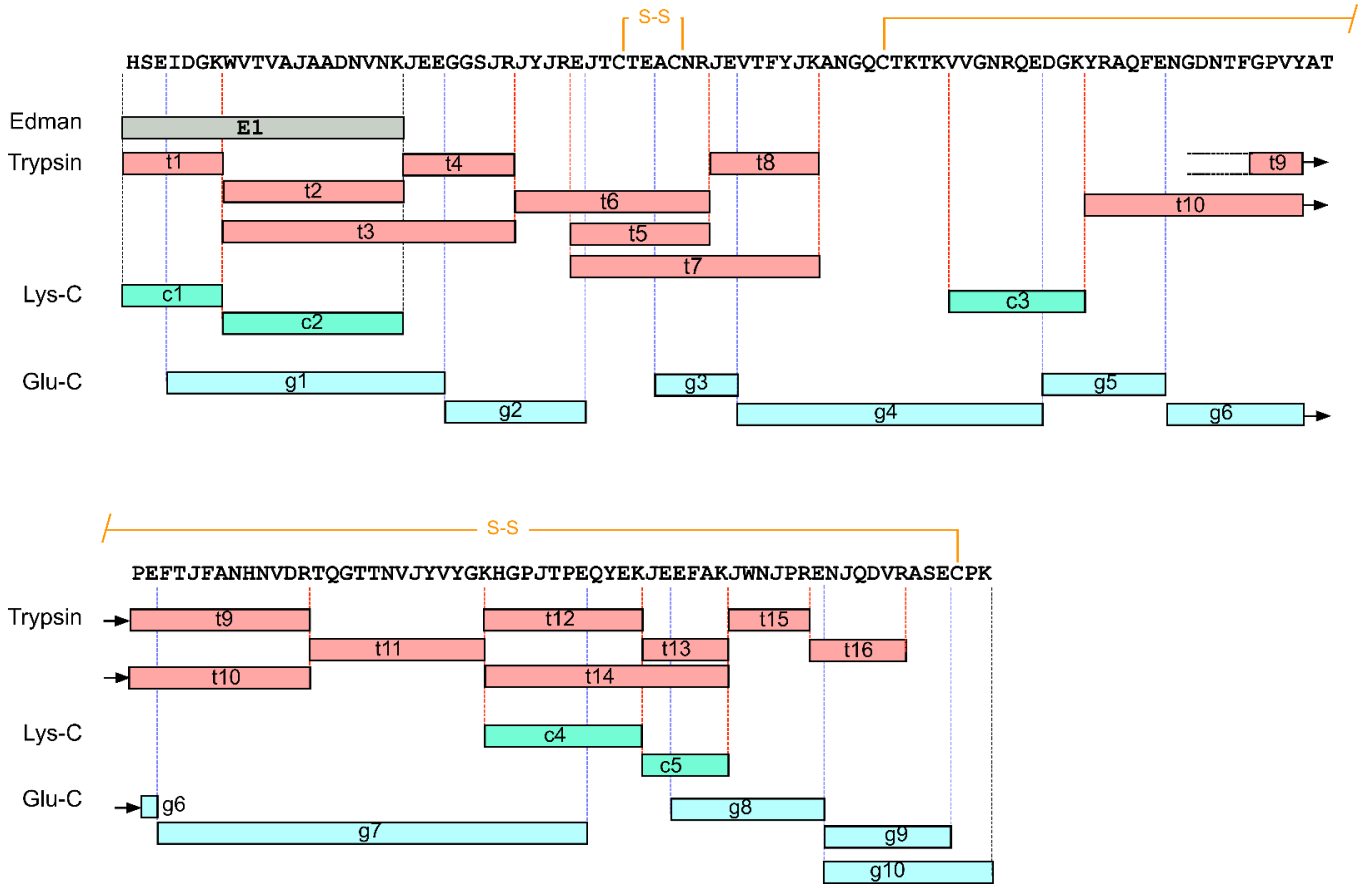


Figure 3.4 | Complete amino sequence of the novel bank vole urinary protein.

The bank vole urinary protein was digested with multiple endopeptidases (t=trypsin, c=endopeptidase LysC, g=endopeptidase GluC) and sequenced de novo by tandem mass spectrometry. In addition, the Edman degradation data of the intact protein allowed definition of the true N-terminus. The symbol 'J' is used to highlight the ambiguity between leucine and isoleucine in all positions other than the N-terminus, where Edman degradation was unambiguous. The positions of the disulfide bond are inferred by homology with similar proteins.

Edman degradation predicted an N-terminal tryptic peptide (HSEIDEK) with a theoretical $[M+H]^+$ mass of *m/z*=857.4. No peptide was detected at $[M+H]^+$ 857.4 Da in either MALDI-TOF MS analysis of trypsin or Lys-C peptides. However, fragmentation of the tryptic peptide at $[M+2H]^{2+}$ *m/z*=393.21 yielded the sequence HSEJDGK (Figure 3.4, peptide t1). This sequence included the highly-conserved glycine residue of the N-terminal lipocalin motif (GXW), aligned with the ambiguous G/E call from the Edman sequencing confirming a glycine residue at this position. The second tryptic peptide within the Edman sequence was predicted as $[M+2H]^{2+}$, *m/z*=700.9; the sequence was determined as WVTVAJAADNVNK (t2) from the b and y ion series using ^{18}O labelling; this contained the tryptophan residue of the GXW conserved motif. The N-terminal region was extended by tandem MS of a miscleaved peptide $[M+3H]^{3+}$, *m/z*=748.08 as WVTVAJAADNVNKJEEGGSJR (t3), also present at $[M+H]^+$ *m/z*=2242.13 in MALDI-ToF MS analysis of tryptic peptides. The sequence of the $[M+H]^+$ 2242.13 tryptic peptide was confirmed by the $[M+2H]^{2+}$ 430.7 tryptic peptide t4, (JEEGGSJR).

Since a feature of OBP-like proteins is the presence of two conserved disulphide bonds, the positions of cysteine residues were identified by carbamidomethylation. MALDI-ToF analysis of tryptic peptides from non-reduced preparations identified two peptides at $[M+H]^+$ *m/z*=1137.51 and *m/z*=2131.04 that were shifted upon carbamidomethylation to $[M+H]^+$ *m/z*=1253.56 and *m/z*=2247.10, a Δmass of 116 Da. The sequence of the reduced and alkylated peptide $[M+H]^+$ *m/z*=1253.56, isolated on LCMS as the $[M+2H]^{2+}$ *m/z*=627.25 (T5) was EJTC*TEAC*NR, containing two modified cysteine residues. The Δmass of 116Da following reduction and alkylation could not be explained simply by the carbamidomethylation of the two cysteine residues which would generate a Δmass of 114.032 Da (2 x 57.016 Da). The additional 2 Da difference is explained by the reduction of a disulfide bond formed between the 2 cysteine residues. Since the unmodified peptide $[M+H]^+$ *m/z*=1137.51 is detected in oxidising conditions, neither cysteine residue could have formed a disulfide bond with a second cysteine residue from a different region of the protein. Further, a high-resolution peptide T6, $[M+2H]^{2+}$, *m/z* = 842.92 sequenced as JYJREJTCTEAC*NR. A tight disulfide loop separated by three amino acids is also a feature of other lipocalins and OBPs, including aphrodisin (Vincent *et al.*, 2001); this provided further presumptive evidence that this protein is an aphrodisin-like lipocalin.

Using similar logic and tandem MS, the entire sequence of the protein was recovered. All high-resolution peptide tandem mass spectra and sequence calls are provided in Supplementary material. The protein sequence predicted a total length of 149 amino acids.

The predicted average mass of the protein was 16934 Da, which, when adjusted to 16930 Da to allow for the loss of 4 Da through formation of the disulphide bonds at C_{36} – C_{40} (proven) and C_{55} – C_{147} (surmised, but consistent with homology modelling) correctly predicted the intact mass measured for the urinary protein.

3.4.3 Discrimination of leucine and isoleucine residues

Mass spectrometry based sequencing *de novo* cannot distinguish between the isobaric amino acids leucine and isoleucine. To discriminate between this isobaric pair, voles were fed a diet partially labelled (relative isotope abundance, RIA, of approximately 0.5) with [$^2H_3$]leucine. Because the protein was secreted in the urine, we surmised that the incorporation of this essential amino acid would result in specific labelling of leucine residues in the protein, and in peptides derived therefrom. Both leucine and isoleucine are essential amino acids, and there is no mammalian metabolic pathway whereby the labelling centres in leucine could be transferred to isoleucine. After digestion, partial labelling meant that each peptide (of monoisotopic mass M) containing a single leucine residue, would be accompanied by a second mass, 3 Da heavier, leading to an M, M+3 Da doublet in both precursor and product ion spectra. Peptides containing solely isoleucine residues would not show any labelling doublet. Lastly, peptides containing more than one leucine/isoleucine residue would require further analysis to locate the position of the leucine residues. The strategy is illustrated in Figure 3.5, together with labelling profiles for several urinary glareosin peptides.

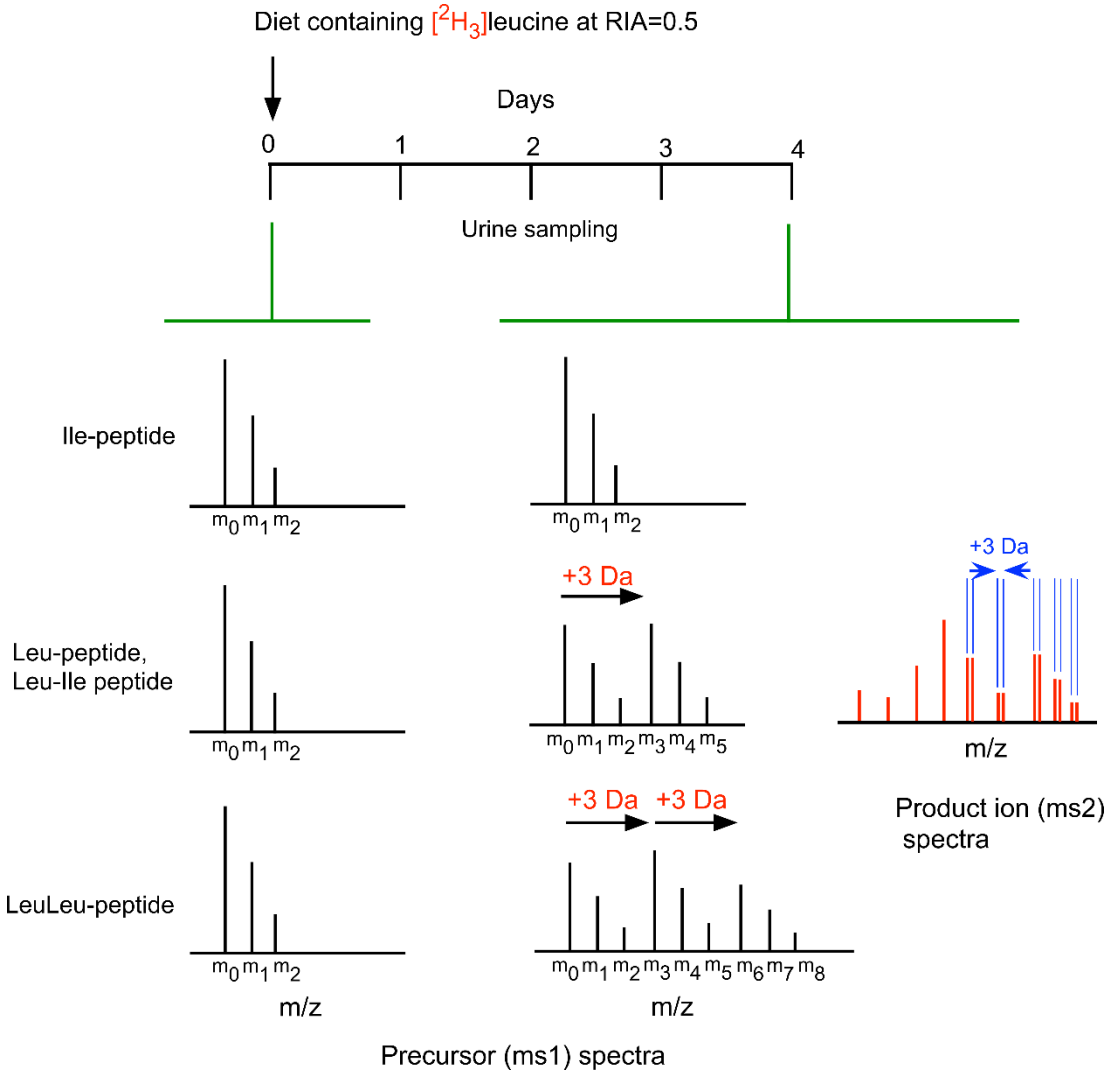
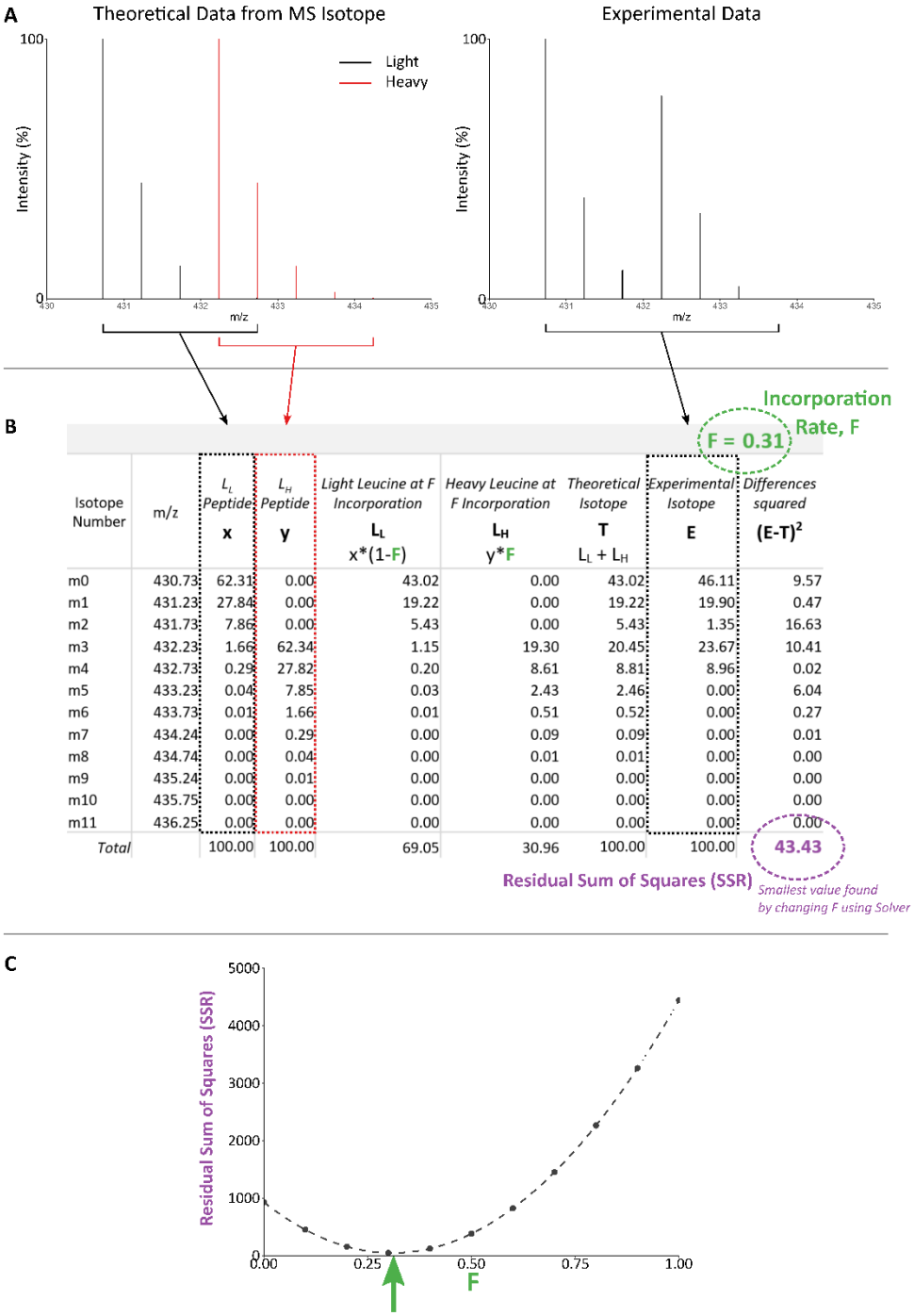


Figure 3.5 | Metabolic labelling strategy to discern leucine and isoleucine residues. Bank voles were fed a diet containing [$^2H_3$]leucine at a relative isotope distribution of 0.5. Incorporation of stable isotope labelled leucine into peptides permits discrimination between leu and ile residues, either from precursor or product ion mass spectra (panel a).

Peptides containing only one ambiguous leucine or isoleucine residue were used to track [$^2H_3$] leucine incorporation over four days of dietary inclusion in four bank voles. The method to calculate incorporation rate used non-linear optimisation and is described in Figure 3.6, as follows. The theoretical isotope intensities for both light and heavy peptides were calculated in MS Isotope (Protein Prospector, UCSF). The proportion of labelled peptide correlates directly with dietary incorporation, so the *m/z* isotope intensities of the **labelled** peptides (L_L) were multiplied by the incorporation factor, F. The proportion of **unlabelled** peptide (L_H) is inversely proportional, so isotope intensities were multiplied by (1-F). Intensities of the labelled and unlabelled peptides for each isotope *m/z* were summed to become an overall **theoretical isotopic spectrum** (T), the distribution of which is

dependent on F. At each isotope's *m/z*, the difference between the theoretical and experimental (E) intensities was calculated and squared. The total of these values, the Residual Sum of Squares (SSR), was then minimised by changing only the value of F using Generalised Reduced Gradient (GRG) non-linear optimisation by utilising the *Solver* function in Excel. The resulting value of F gave the best-fitting incorporation rate for the data observed, and this process was repeated for three peptides with consistently high quality data for all four voles over the four day dietary incorporation period (Figure 3.7). By calculating the incorporation rate for each vole on each day sampled, it was possible to confirm that the dietary isotope labelling was correctly applied, by observing an incorporation rate close to 50%, which plateaued over the feeding time.



Isotope Number	m/z	L_L Peptide x	L_H Peptide y	Light Leucine at F Incorporation L_L x*(1-F)	Heavy Leucine at F Incorporation L_H y*F	Theoretical Isotope T $L_L + L_H$	Experimental Isotope E	Differences squared $(E-T)^2$
m0	430.73	62.31	0.00	43.02	0.00	43.02	46.11	9.57
m1	431.23	27.84	0.00	19.22	0.00	19.22	19.90	0.47
m2	431.73	7.86	0.00	5.43	0.00	5.43	1.35	16.63
m3	432.23	1.66	62.34	1.15	19.30	20.45	23.67	10.41
m4	432.73	0.25	27.82	0.20	8.61	8.81	8.96	0.02
m5	433.23	0.04	7.85	0.03	2.43	2.46	0.00	6.04
m6	433.73	0.01	1.66	0.01	0.51	0.52	0.00	0.27
m7	434.24	0.00	0.29	0.00	0.09	0.09	0.00	0.01
m8	434.74	0.00	0.04	0.00	0.01	0.01	0.00	0.00
m9	435.24	0.00	0.01	0.00	0.00	0.00	0.00	0.00
m10	435.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00
m11	436.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total		100.00	100.00	69.05	30.96	100.00	100.00	43.43

Figure 3.6 | Determining the incorporation rate of heavy leucine-labelled peptides. Theoretical intensity ratios for mass isotopes from both light (black) and heavy (red) peptides were generated in MS-Isotope and compared to the experimental isotope distribution of that peptide (**A**). The theoretically-generated intensity isotopes of the labelled peptide (red; L_H), 'y' were multiplied by the incorporation factor, F, and the theroetial isotopes of the unlabelled peptide (black), 'x', were multiplied by (1-F). The sum of the relative intensity for each isotope *m/z* was therefore the total theoretical isotope distribution at a given incorporation rate. The differences between the theoretical distribution and experimental data were squared and summed. This Residual Sum of Squares (SSR) was minimised by changing only the incorporation factor, F, using the Excel function Solver, an algorithm for non-linear optimisation using Generalised Reduced Gradient (GRG) (**B**). A visual representation of this method is illustrated in **C**. The SSR value for this particular example is plotted against seeded

values of F (0.0-1.0). The algorithm finds the lowest possible SSR value. The closest estimation of the incorporation value, F, in this example, was 0.31 (green arrow).

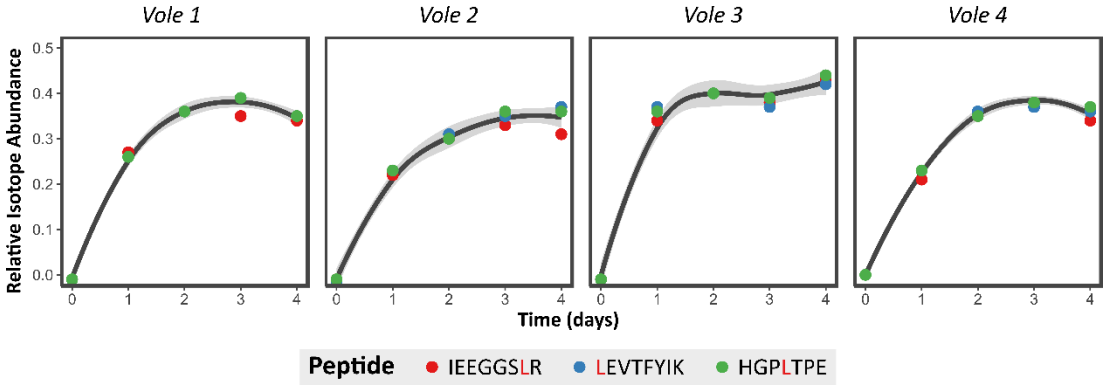


Figure 3.7 | *Dietary incorporation of 5,5,5-d3 leucine over time*.
After four days of labelling, urinary proteins (three representative peptides are shown here for four bank voles) were labelled to the same extent as the dietary precursor and used for sequencing *de novo*.

All leucine/isoleucine ambiguities were evaluated manually and assigned from the raw data. Tryptic peptides containing a single ambiguous site (defined as 'J'; HSEIDGK isobaric identity known from Edman degradation, WVTVAJAADNVNK, EJTCTEACNR, TQGTTNVJYVYGK, HGPJTPEQYEK, ENJQDVR, ACNRJE, VTFYJK, FTJFANHNVDR) were readily resolved from the precursor ion spectra. For peptides that contained more than a single instance of leucine or isoleucine, the strategy was more complicated. Most simply, precursor ion spectra could disambiguate peptides that contained two of the same residues (JWNJPR), as the mass shift was unambiguous (+ 6 Da, Leu/Leu; 0 Da: Ile/Ile). For peptides that contained two Leu/Ile residues, only one of which was labelled, the precursor ion mass shift indicated the number but not the position of the leucine and isoleucine residues. Positional resolution was achieved by inspection of fragment ion spectra. Fragment ion spectra were examined for +3 Da increases in the y- and b-ion series at each Leu/Ile product ion, clearly flagged as a doublet. This defined the position of Leu and Ile for peptides JEEGGSJR (Figure 3.8), JEVTFYJK and JYJRE (Supplementary).

For each of these three peptides, fragment ion data for all four voles were compared and confirmed the same heavy leucine sites. Examples of the fragment ion data from the heavy-labelled parent ion, $[(M+3L_H)+n]^{n+}$, of each peptide are found in supplementary.

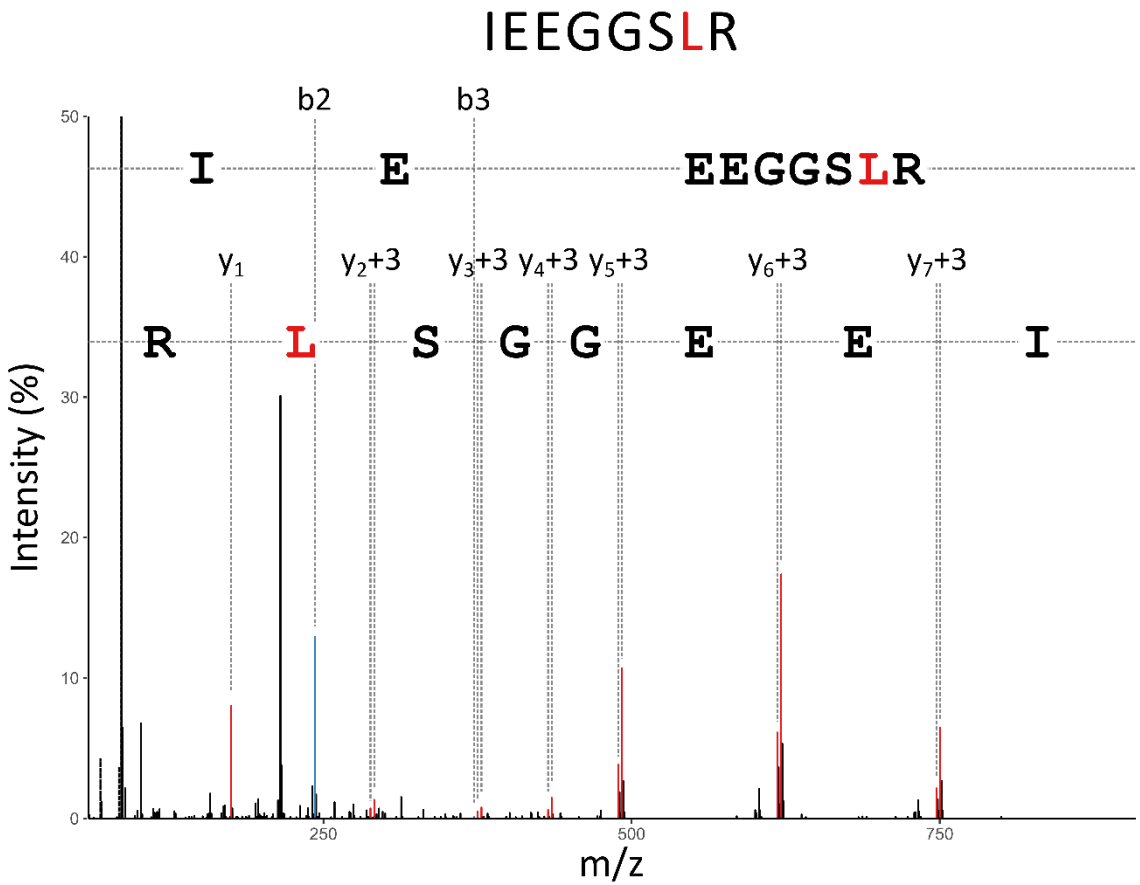


Figure 3.8 | Determining leucine and isoleucine residues from fragment ion spectra.
Ambiguous leucine and isoleucine residues that could not be determined by MS1 spectra were established using product ion spectra. The product ion spectrum for the peptide IEEGGSLR is shown above as an example, from the heavy labelled parent ion ($[M+2H]^{2+}$, 432.3 m/z]). A Δ3 Da is observed in the y-ion series from the y_2 ion onwards (denoted y_n+3), showing the placement of the heavy labelled leucine. Example spectra for the remaining peptides can be found in supplementary material.

The remaining unassigned Leu/Ile site was in the small tryptic peptide JEEFAK (t13; $[M+H]^+$= *m/z* 736.3875), which is identical to an equivalent tryptic peptide derived from OBP2 and OBP3. To resolve this issue, we assigned the residue identity using tryptic missed cleavage peptides (this work: HGPLTPEQYEKJEEFAK compared to the peptide GQPLTPEQYEKLEEFAK from OBP2 and OBP3 (Uniprot D3VW63_MYOGA and D3VW64_MYOGA), respectively. The first leucine residue for the protein described here had already been confirmed (see previously).

Therefore, the peptide could either be HGPLTPEQYEKLEEFAK or HGPLTPEQYEKIEEFAK. To resolve this, the **experimental** isotope distribution (E) of the both potential missed cleavage peptides from glareosin, and the missed cleavage peptide from the OBPs (GQPLTPEQYEKLEEFAK)were compared to their **theoretical** distributions (T), given from the combination of heavy and light peptide *at the incorporation rate (F) determined previously* for each vole, as described in Figure 3.6.

Firstly, the fit of the experimental data was compared to a theoretical isotope distribution of a Leu-Ile peptide, given an incorporation factor of approximately 0.4 (exact value dependent on individual animal), based on the method described above (see Figure 3.6). Secondly, data were compared to a theoretical distribution of a Leu-Leu peptide at 40% incorporation. However, effect of F differed when considering mass increments of both +3 and +6 Da. The proportion of the peptide where both ambiguous sites are labelled is no longer directly proportional to the incorporation rate, but equal to $(F)^2$. Similarly, the proportion of unlabelled peptide is given by $(1-F)^2$. The proportion of peptides containing one labelled residue will increase to a maximum at 50% incorporation but reduce with further incorporation. Thus, the proportion of the Leu-Ile peptide is described by the equation $y = -2(F)^2 + 2(F)$ (Figure 3.9). The theoretical isotope distribution for a Leu-Leu peptide at a given incorporation rate was therefore calculated.

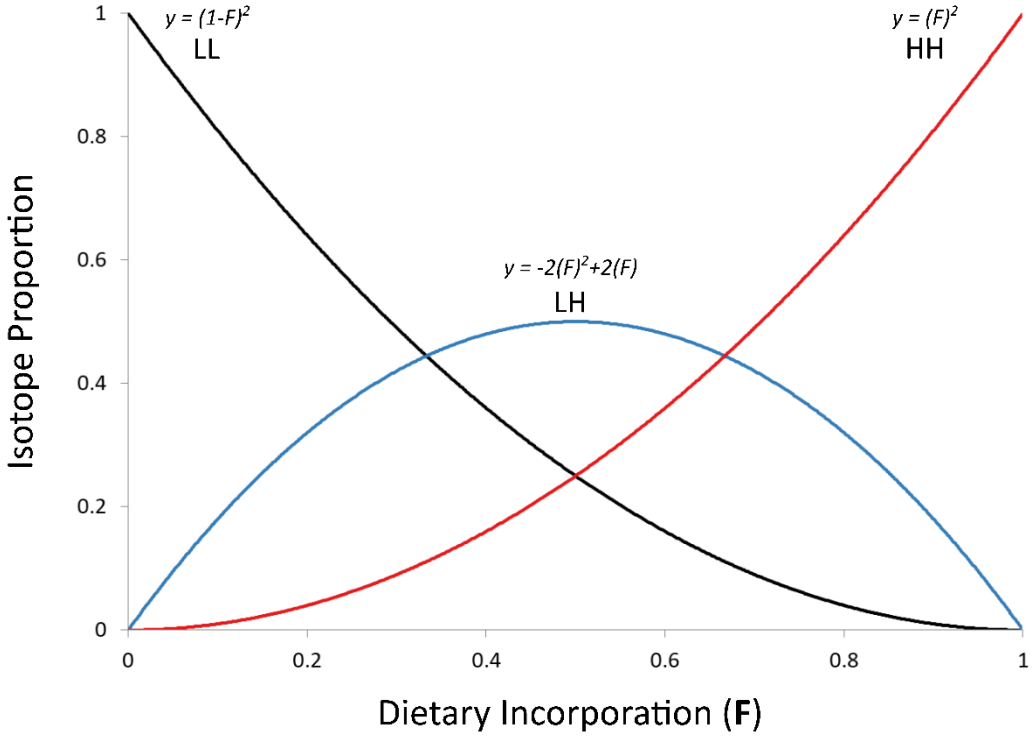


Figure 3.9 | Proportion of differentially labelled peptides during 5,5,5-d3 leucine incorporation with two ambiguous leu/ile sites.
The proportion of unlabelled peptide (L_L- L_L), peptide with a single labelled leucine site (L_H-L_L) and fully labelled peptide (L_H-L_H), are displayed in black, blue and red, respectively, over incorporation.

The SSR values, representative of the fit of the **observed** isotope distribution for the glareosin peptide to the two **theoretical** potential peptide sequences, containing either one or two labelled leucine sites, were calculated using the method described above, with the modified equations to accommodate two ambiguous sites. The same method was also applied to the OBP peptide data. The SSR value for each experimental-theoretical comparison was compared, and in all animals, the best fit for the glareosin peptide was one leucine (previously determined), one isoleucine, with total squared differences far lower compared to a theoretical distribution considering two labelled sites. In contrast, the OBP peptide fitted a distribution distinctive of a peptide with two labelled leucine residues. An example comparison is described in Figure 3.10, using experimental data from Vole 4 on day 4 of incorporation.

The precursor isotope masses of the missed cleavage peptide had a distribution consonant with one leucine residue and one isoleucine residue, whereas the OBP peptide also present

in LC-MS/MS analysis displayed a fragment ion distribution consistent with the presence of two heavy leucine residues (Figure 3.10).

All ambiguous leucine/isoleucine residues could therefore resolved (Figure 3.11).

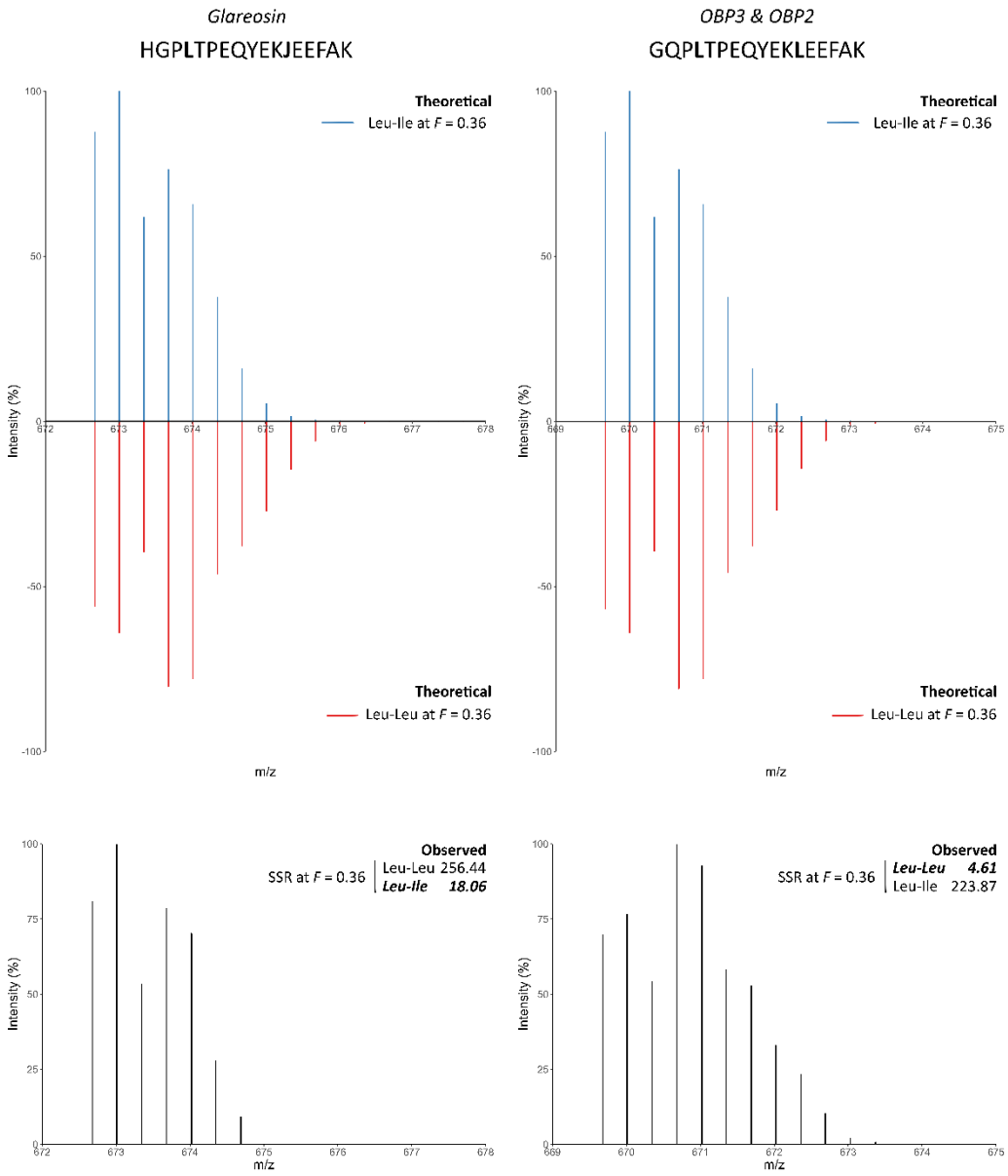


Figure 3.10 | *Resolution of leucine/isoleucine residues in a missed cleavage peptide.*
The isotope distribution from the missed cleavage tryptic peptides in glareosin (left panels) and OBP3 & 2 (right panels) were used to determine ambiguous leucine and isoleucine sites. The theoretical distribution for peptides with two labelled leucine sites (red) and one labelled leucine site (blue), at an incorporation of 36% (calculated previously for this animal, vole 4, on day 4 incorporation) were generated (top panels). The sum of the differences squared (see previous) was calculated by the comparison of experimental data to each theoretical spectrum and the ambiguous site in the glareosin peptide was determined as an isoleucine.

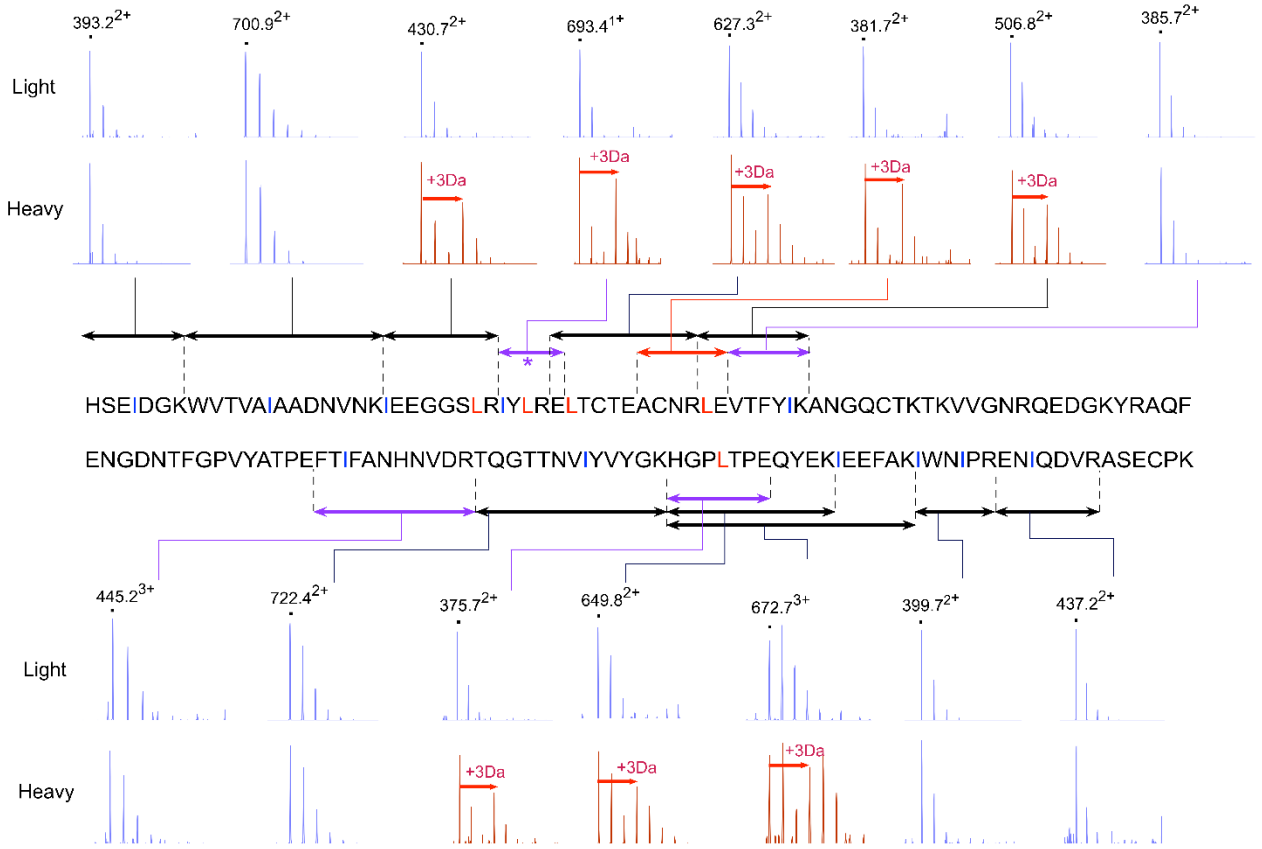


Figure 3.11 | Resolution of leucine and isoleucine by metabolic labelling.

After dietary administration of [$^{2}H_{3}$]leucine, proteolysis and mass spectrometry of the bank vole lipocalin, the assignment of leucine and isoleucine residues was completed. The figure indicates the residue assignment annotated with the precursor mass spectrum of the appropriate peptide (double headed arrows), generated from trypsin (black), endopeptidase GluC (red) or digests using both endopeptidases (purple). The monoisotopic, unlabelled ion is marked with a black dot. Spectra that confirm the incorporation of a stable isotope labelled leucine residue are coloured red and the mass offset (3 Da, due to incorporation of labelled leucine) is indicated with a red arrow.

3.4.4 *Phylogenetic analysis*

We were thus able to derive the complete, unambiguous sequence of the bank vole urinary protein, including the identification of all leucine and isoleucine residues. The entire sequence was used in a BLAST search against all rodent sequences. The first major conclusion is that this abundant protein in bank vole urine is novel, and has not been reported previously. To distinguish this protein from other bank vole urinary proteins (Stopková *et al.*, 2010a) we therefore propose the name 'glareosin' (derived from the species *Myodes glareolus*). The glareosin sequence matched to several lipocalins, most strongly to aphrodisins and odorant binding proteins (OBPs), with weaker matches to probasins (prostate expressed 'outlier' lipocalins) and MUPs. A phylogenetic tree (Figure 3.12) defines the relationships between these groups of lipocalins, specifically those from rodents and a full alignment is given in Supplementary material.

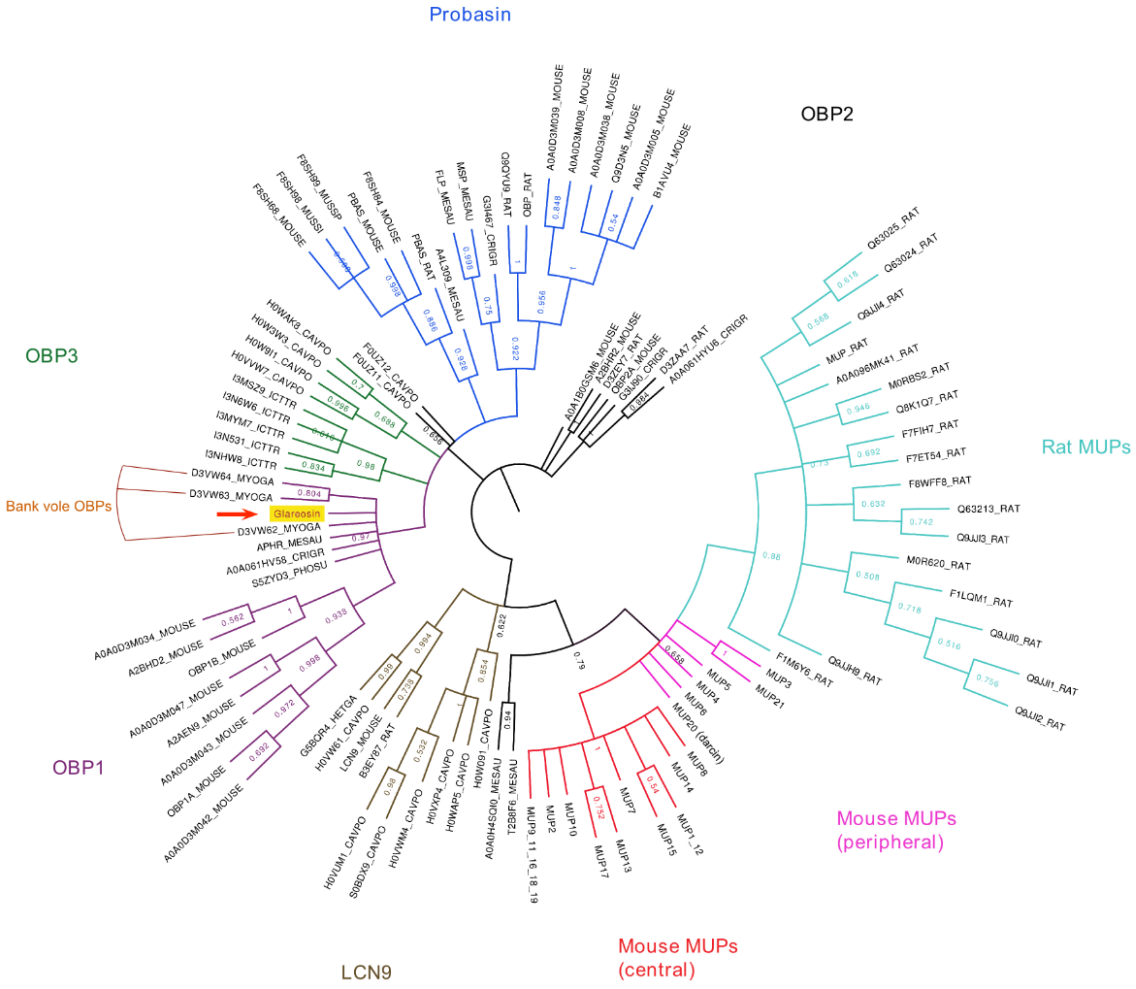


Figure 3.12 | *Phylogenetic tree of glareosin-related sequences.*
Bootstrapped Maximum Likelihood phylogenetic tree calculated using MEGA7 as described in Methods. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. With the exception of a manually curated set of mouse MUPs based on the MGI database

(http://www.informatics.jax.org/searchtool/Search.do?query=mup*), proteins are labelled with UniProt identifiers. The three OBPs previously identified in bank voles (Stopková *et al.*, 2010a) are highlighted.

Of interest is the relationship between glareosin and the odorant binding proteins (OBPs 1a, 1b, 2 and 3) that have previously been detected in bank vole urine samples (Stopková *et al.*, 2010a). The four proteins share over 60% sequence identity, and the presence of the lipocalin GXW motif and the disposition of the two disulphide bonds means that all four proteins share a high level of structural and possibly functional similarity. Yet, glareosin was not discovered or described as the predominant urinary protein in the previous study (Stopková *et al.*, 2010a), in which the two urinary proteins detected on two-dimensional electrophoresis followed by mass spectrometry were OBP2 and OBP3. Indeed, when we perform a discovery proteomics analysis on a tryptic digest of total urinary proteins, we also see good evidence for these two proteins in bank vole urine (data not shown) but at a much lower level than peptides derived from glareosin. On a one-dimensional SDS-PAGE gel, glareosin is by far the most strongly expressed protein, and at first glance, it is not obvious why this protein was not observed in the previous study. However, analysis of the sequence of glareosin and the three OBPs reveals that the predicted isoelectric point (pI) of OBPs 1 to 3 are 5.0, 4.8 and 4.8 respectively. By contrast, the predicted pI of glareosin is 5.7. In the previous study (Stopková *et al.*, 2010a), the pI range of the two-dimensional gel system used to visualize and identify urinary proteins was from 3.9 to 5.1. It is highly likely that glareosin was not resolved by the first, isoelectric focusing dimension, would not have entered the gel and thus could not have been detected.

3.4.5 Structural homology modelling

The complete protein sequence derived by mass spectrometry, including disambiguation of leucine/isoleucine, allowed us to submit the primary sequence to three-dimensional structure prediction. Of predicted structures for glareosin, those produced with a single alignment to aphrodisin (58–60) consistently scored better than those produced with either the top five or ten templates identified by HHpred. Aphrodisin is distinctly more closely related to glareosin (47% sequence identity) than other templates (39% at most) - this reinforces the observation that inclusion of more distantly related templates does not always benefit model quality when a closely homologous structure is available. Models generated with the initial HHpred alignment of glareosin with aphrodisin consistently exhibited stereochemical problems near the C-terminus where glareosin has a one-residue deletion compared to aphrodisin. Examination of the aphrodisin structure suggested that side chain interactions would be better retained with a one-residue shift of the deletion

position. Positioning the deletion opposite Thr_{149} (mature protein sequence) in the aphrodisin template eliminated serious stereochemical issues and produced better scoring models by validation metrics.

Unexpectedly, the final model set contained two distinct conformations which scored equally well by all criteria. Each conformation gives a normalised QMEAN Z-score of 0.44 showing that the structures, by the six distinct component scores considered, perform slightly better than the average protein of a similar size. The two conformations differ in the position of loop 5, connecting ß-strands E and F (in the standard family nomenclature (Flower, 1996)). In the 'closed' conformation, the loop lies over the entry to the central binding pocket, as is typically observed in crystal structures (Figure 3.13a), while in the 'open' conformation the entrance to the binding pocket is unimpeded and the pocket connects directly with bulk solvent. The validity of the two conformations is supported by the ability of the Rosetta program to accurately sample alternative, biologically relevant conformations: it has proved capable of predicting a second allosteric state accurately, given a crystal structure of the first (Kidd, Baker and Thomas, 2009). Pathways of interconversion between these two conformational states could be explored in the future by Molecular Dynamics simulations. Interestingly, this loop bears a unique one residue insertion compared to all near relatives of known structure. Thus, it is possible that glareosin has distinct ligand binding properties when compared to other semiochemical lipocalins whose crystal structures, with cavity occupied or empty, show a strong tendency towards closed structures (Figure 3.13b).

The central, beta-barrel enclosed cavity of glareosin has a similar volume to aphrodisin; GHECOM (47) estimates them as 305 $Å^3$ and 318 $Å^3$, respectively, while the volumes from Profunc are 357 $Å^3$ and 377 $Å^3$. The cavity of the model structure of glareosin is more elongated, hinting at possible differences in specificity of bound ligands (Figure 3.13c). For comparison, GHECOM predicts a cavity of 324 $Å^3$ and Profunc, 410 $Å^3$ for the unoccupied MUP (1I04.PDB) and GHECOM: 396 $Å^3$. Profunc:450 $Å^3$ for a cavity occupied MUP (1I04.PDB). The glareosin cavity is thus of lower volume than the MUP but is still large enough to accommodate a broad range of low molecular weight ligands. Ligands of glareosin have yet to be identified.

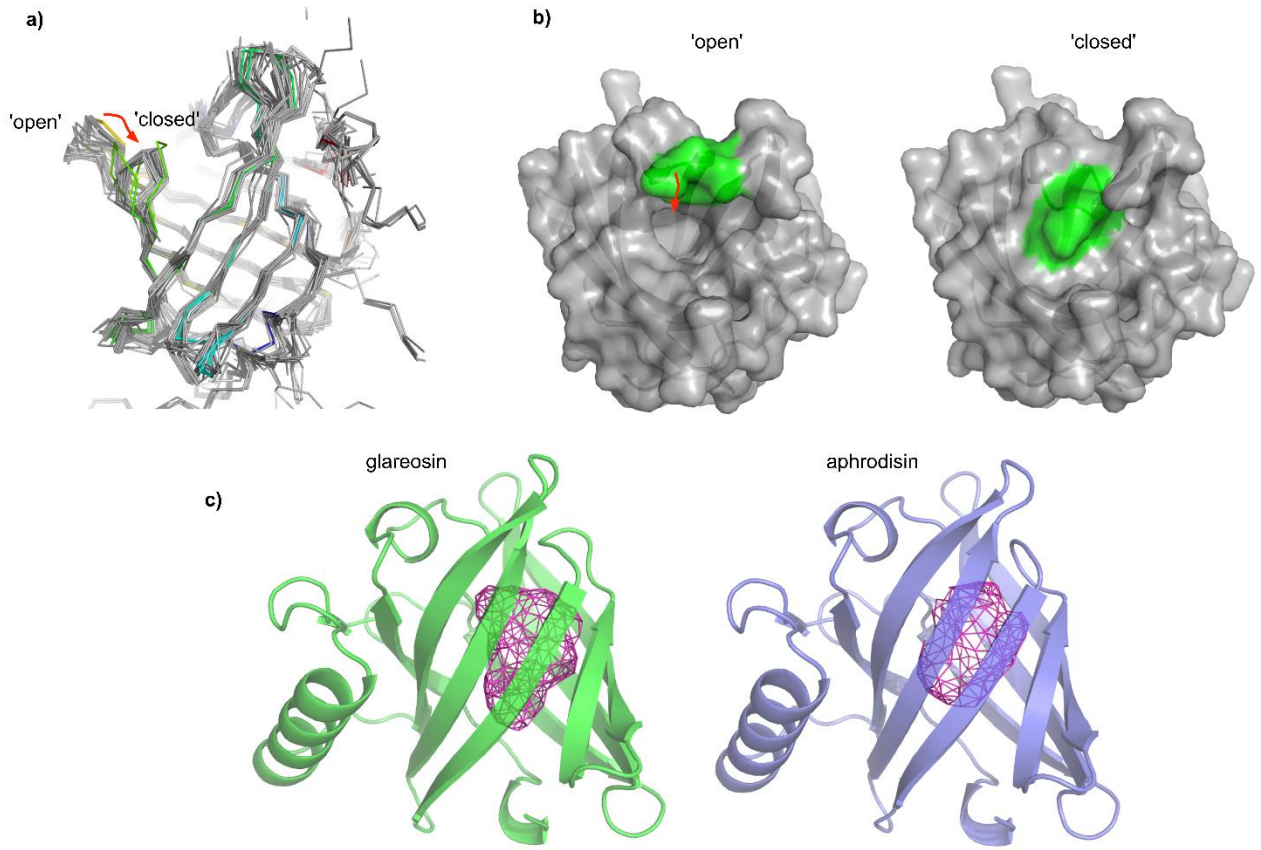


Figure 3.13 | ***Predicted three-dimensional structure of glareosin.***
The structure of glareosin was predicted by homology modelling. Two solutions (an 'open' and a 'closed' conformation) were predicted equally well (panel a and b). In a) the two solutions are coloured blue to red from N- to C-terminus with all experimental structures of lipocalins sharing at least 25% sequence identity with glareosin shown in grey. In b) the loop differing in conformation is shown as green, the rest of the glareosin models as grey. The cavity at the centre of the closed glareosin structure was analysed using the Profunc server (Laskowski, Watson and Thornton, 2005) and compared to aphrodisin (panel c).

It has previously been reported that the urinary protein output of *M. glareolus* is sexually dimorphic and that males exhibited obligate proteinuria in all sample types investigated (Kruzcek et al., 1985). Males mark new territory in frequent small drops without entirely emptying their bladder, compared to females that deposit large pools of urine (Johnson., 1975 Christiansen., 1980). This sex specific behaviour is similar to that of the house mouse, where the repeated marking of territory with small volumes of urine is used to advertise competitive dominance (Desjardins, Maruniak and Bronson, 1973; Hurst and Beynon, 2004).

3.4.6 Further identification of proteins in bank vole urine

The urinary protein profile of bank voles, a seasonal, sex-specific expression of a single predominant protein, appears to be relatively simple. However, global investigation into protein expression could reveal if there are less abundant proteins that could still be contributing to a signalling profile.

As an initial exploration, proteins from pooled male and pooled female bank vole urine were separated by SDS and native PAGE (10 µL each) and resulting gel bands were excised and subject to protease digestion. The resulting peptides were analysed by LC-MS/MS and identified in PEAKS^{TM} by searching against a database of all *Rodentia* protein sequences in UniProt (Figure 3.14). In most cases, the top-scoring gel band, after discarding trypsin autolysis products and keratin contamination, was used to identify the gel band. In some cases, the top-scoring protein identified was not likely to be correct, for example due to contamination from neighbouring protein bands, which was particularly prominent from native PAGE separation. Consequently, other parameters were used to assist protein identification, including the number of peptide-spectrum matches made, the number of peptides identified, and total area under the spectral curve (AUC) for those peptides. In addition, the peptide candidates generated *de novo* by PEAKS^{TM} for unmatched spectra were inspected to assess if other protein candidates could be manually identified (see sections 1.5.2 and 1.5.3 for more information on the strategy employed by PEAKS^{TM}).

One of the issues encountered was multiple identification for the **same** protein entered into the database, but for multiple species. Protein grouping in proteomics software tackles this issue to some degree, by grouping together proteins that are identified by the same set of peptides. However, due to the nature of cross-species matching, some observed peptides will match a segment of a homologous protein, or proteins, whilst other may

match another section. If the database used is compiled of protein sequences from multiple organisms to increase the chances of identifying a peptide, these sections might be identified from the same protein from different species, and therefore not form a peptide subset that could be recognised as a singular protein group by the software, but as two separate entities. Not only does this result in multiple identifications that makes a direct comparison difficult, but also reduces the number of unique peptides, a parameter which is commonly used to filter out lower-confidence identifications. For example, peptide data from the tryptic digest of the slowest resolving band from SDS-PAGE analysis of pooled male bank vole urine identified seven uromodulin sequences from four species, forming four groups (Table 3.1). Therefore the most likely candidates for in-gel identifications were manually curated. The total protein score, total area under curve, number of peptides and scans for each protein group were considered, in addition to the typical mass of homologous proteins.

Table 3.1 | Cross-species matching with multiple species databases results in multiple identifications of the same protein.
Proteins from pooled male and female bank vole urine were separated by SDS and native PAGE. The resulting protein bands were excised and subject to tryptic digestion prior to LC-MS/MS analysis, the data from which was analysed in PEAKS™. The resulting identification of the slowest-resolving protein band from SDS-PAGE analysis of pooled male bank vole urine was uromodulin, and seven sequences from four species were identified. Protein grouping results in four sets of proteins.

PROTEIN GROUP	ACCESSION	-10LOGP	COVERAGE (%)	#PEPTIDES	#UNIQUE	AVG. MASS	DESCRIPTION	SPECIES
1	G3ILJ3	207	16	13	5	67809	Uromodulin	*Cricetulus griseus*
1	A0A061I9T1	207	8	13	5	137390	Uromodulin-like protein	*Cricetulus griseus*
4	A0A0P6JXY0	193	10	8	2	70083	Uromodulin	*Heterocephalus glaber*
4	G5C2G5	193	10	8	2	72302	Uromodulin	*Heterocephalus glaber*
2	Q91X17	171	15	10	4	70845	Uromodulin	*Mus musculus*
3	A0A0G2JSP1	161	11	7	2	71016	Uromodulin	*Rattus norvegicus*
3	P27590	161	11	7	2	71062	Uromodulin	*Rattus norvegicus*

The most confident identification from cross-species matching was albumin; in both types of gel this was the third slowest resolving protein band. It was the highest-scoring protein in all four datasets, with the highest-10lgP scores for each at 273, 258, 324 and 190 for male (SDS), female (SDS), male (native) and female (native), respectively.

The slowest resolving protein band in both SDS and native PAGE, for both males and females, was identified as uromodulin, andthe protein band resolving between albumin and uromodulin in both native and SDs-PAGE was identified as serotransferrin.

Albumin, uromodulin and serotransferrin are all proteins commonly seen in urine. Uromodulin is produced in the loop of Henle (Bachmann, Koeppen-Hagemann and Kriz, 1985; Bachmann, Metzger and Bunnemann, 1990), is commonly found in mammalian urine (Serafini-Cessi, Malagolini and Cavallone, 2003) and may play a regulatory role in the kidney (Bachmann *et al.*, 2005). Serotransferrin is a serum glycoprotein with iron-binding capacity that has been identified in urine (Chasteen, 1977; Wait *et al.*, 2001). Albumin is the most common protein in blood plasma and elevated abundance in urine is a common biomarker for renal dysfunction (Anderson and Anderson, 2002; Martin, 2011). The presence of each of these proteins is unsurprising, and unlikely to be relevant to any semiochemical function.

The remaining, lower molecular weight bands were identified from previously established sequences for glareosin and odorant binding proteins (Stopková *et al.*, 2010a). Protein bands from male urine corresponding to glareosin were easily identified, with -10lgP scores of 228 and 210 from SDS and native PAGE bands, respectively. OBP3 was the highest scoring protein in the fourth fastest running band from SDS-PAGE separation. The fifth fastest resolving gel band from native PAGE separation of male bank vole urine was also deemed most likely to be OBP3, as it was the second-highest scoring protein. The highest scoring protein was glareosin, but the total area under spectral curve for OBP3 was higher, so this was considered contamination from the neighbouring protein band.

Three other proteins in each SDS and native PAGE separation of male bank vole urine could not be determined, as glareosin, OBP3 and OBP2 were all identified as potential candidates. Manual inspection of unmatched, high quality, abundant sequences generated *de novo* by the software also suggested potential peptides with homology to OBPs and glareosin, but did not generate an identification for any established sequences. An example is data from the gel band at approximately 25 kDa from the SDS-PAGE gel, from which 7 glareosin peptides, 8 OBP3 peptides, 4 OBP2 peptides and a further 10 homologous manually-identified peptides were discovered (Figure 3.15). Consequently, no definitive identification

was made for these protein bands and they have been deemed 'OBP-like'. A sequence alignment of glareosin and the bank vole OBPs was constructed, together with peptides identified from in-gel digestion for these sequences, and compared to peptides sequenced *de novo* by PEAKS™ from unidentified spectra and manually identified as OBP-like (Figure 3.15). This gives an example of the peptide-level heterogeneity observed in experimental data that is yet to be explained.

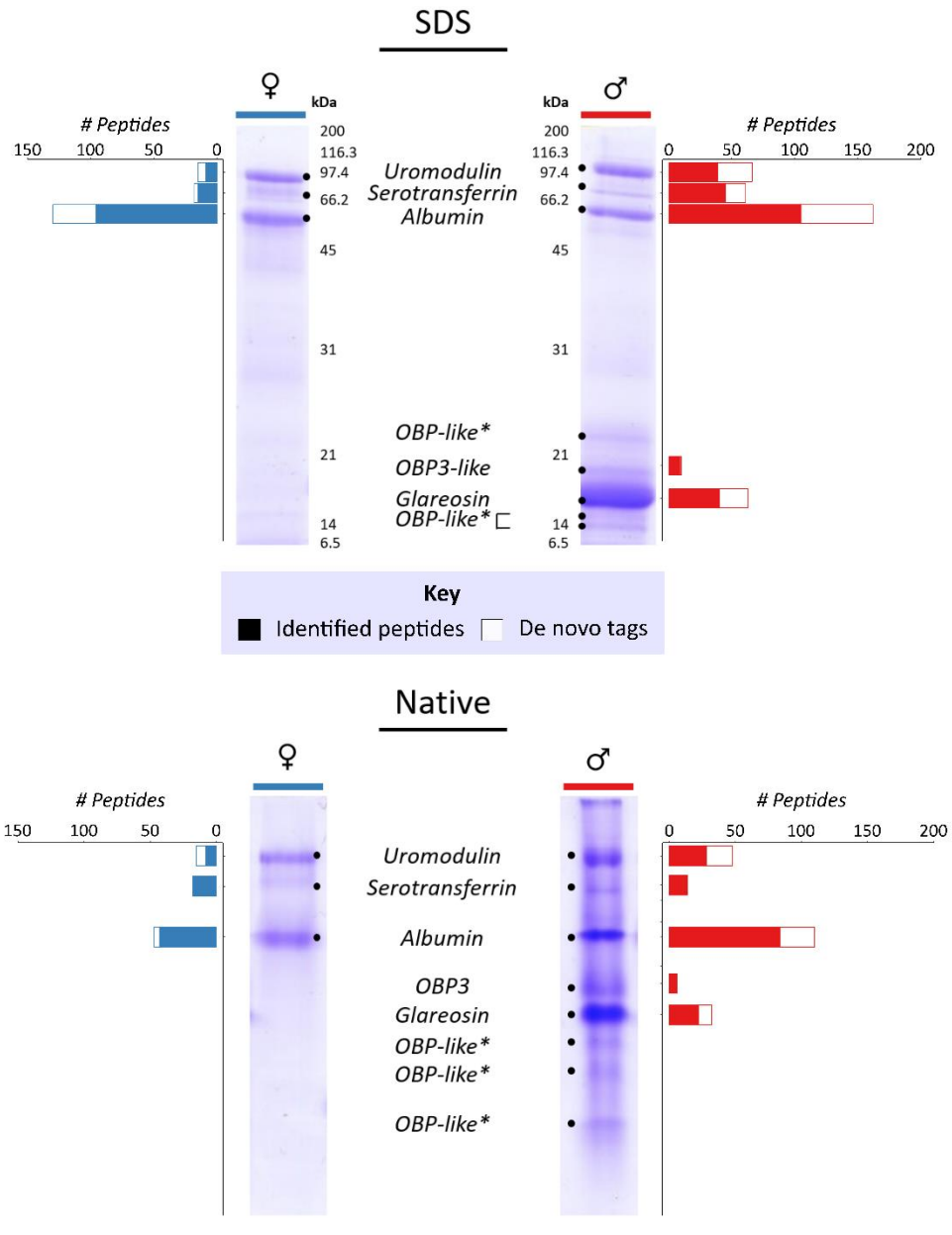


Figure 3.14 | In-gel identification of major protein bands separated by SDS and native PAGE. Proteins in pooled male and female were separated by SDS and native PAGE. Separated proteins were digested in-gel and analysed by LC-MS/MS. Proteins were identified in PEAKS™ by searching against a database of available *Rodentia* sequences from UniProt. Identification was based on top-scoring proteins combined with manual confirmation from PEAKS™-generated *de novo* sequenced peptides. * = specific protein not confirmed.

Whilst it is clear that glareosin is the most abundant protein in male bank vole urine, and dominates both intact mass profiles and peptide data, in-gel digestion revealed protein complexity at a more subtle level. Identification of OBP3 and OBP2 sequences indicates that the proteins identified by Stopková et al. (2010a) may be present in this cohort of bank vole urine samples, although what remains to be established is the continuity of the primary protein sequences. It would be unsurprising if the OBP genes that were established from an eastern European bank vole population contained some sequences differences when compared to those observed in bank vole urine from a UK animal.

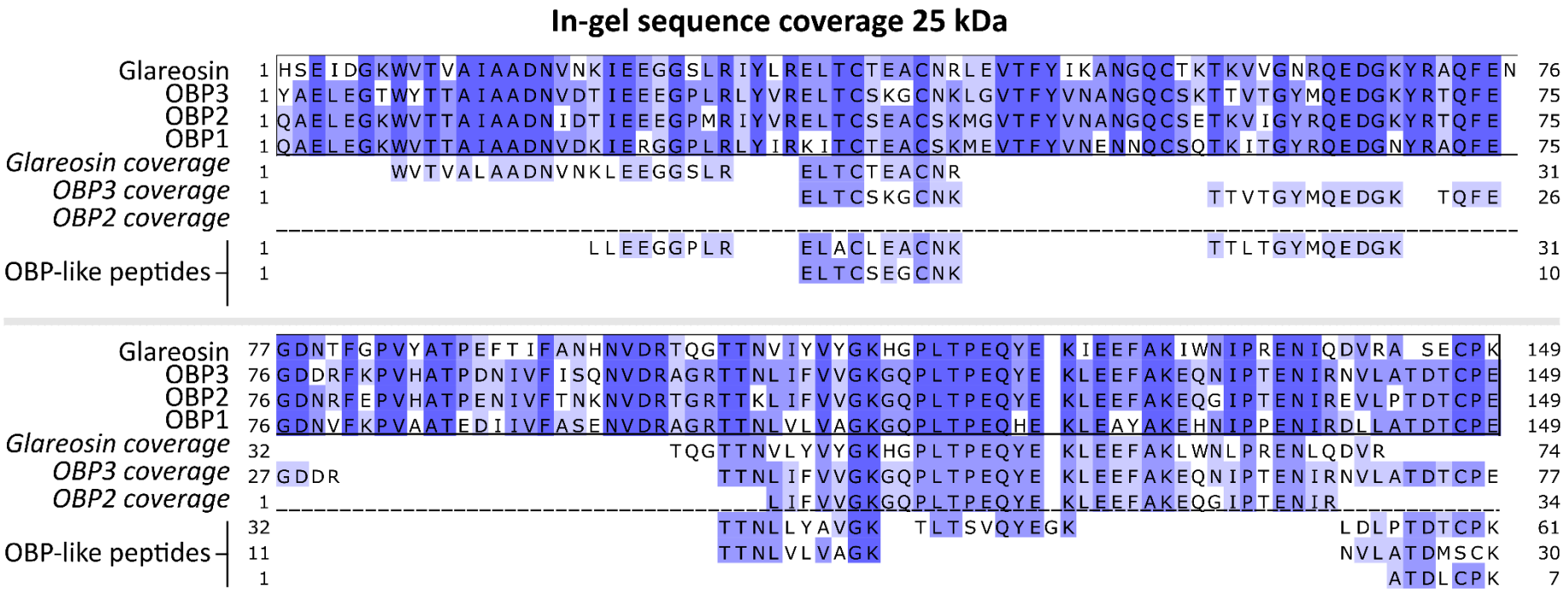


Figure 3.15 | In-gel identification of major protein bands separated by SDS and native PAGE.
Multiple OBP-like peptides were mapped to several bands, and distinction between the candidate proteins was not possible. Example shown is sequence coverage observed from 25 kDa SDS-PAGE protein band. Signal peptides were first removed from the odorant binding sequences from UniProt using the SignalP server (Petersen *et al.*, 2011). The remaining sequences were combined with the glareosin sequence, the peptide coverage observed from in-gel identification for glareosin, OBP3 and OBP2, and the manually identified OBP-like peptides and aligned using Clustal Omega (Sievers *et al.*, 2011). The alignment was reviewed and formatted in JalView (Waterhouse *et al.*, 2009).

The identifications made from pooled bank vole urine were confirmed with in-gel data from individual urine samples. Native PAGE of seven male bank vole urine samples was used to separate proteins for in-gel digestion and analysis by LC-MS/MS. The same rationale was taken to identify the main protein(s) contributing to each protein band (Figure 3.16). Again, three gel bands, which were only visible in one sample, were recognised as homologous to glareosin and the previously reported OBPs, but homologous sequences were also manually identified.

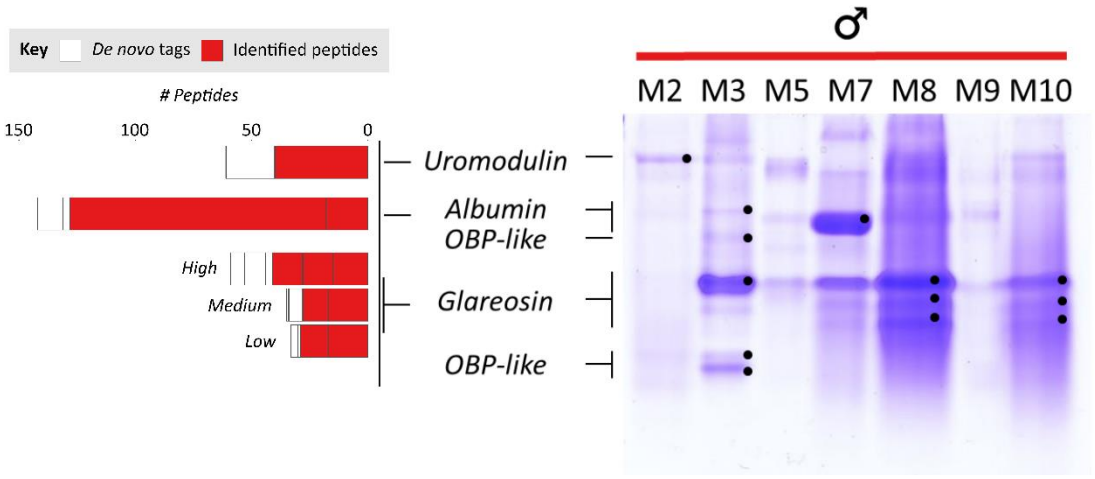


Figure 3.16 | In-gel identification of major protein bands separated by native PAGE. Identifications from in-gel digestion of pooled male urine were confirmed from proteolytic digestion of proteins separated from individual samples (n=7).

In male bank vole urine, glareosin is the overwhelmingly dominant protein. However, in-gel proteolytic digestion of proteins separated by SDS and native PAGE and analysed by LC-MS/MS retrieved peptide-level evidence for the presence of odorant-binding proteins seen previously (Stopková *et al.*, 2010a), in addition to homologous peptides from unidentified *de novo* sequences. However, it was not possible to gather enough information from in-gel digestion to assess the number of novel OBP-like proteins present in bank vole urine. Nor was it possible to assess if the known OBP sequences identified were sequentially fully identical to those reported by Stopková et al. (2010a) because full sequence coverage was not achieved. Ideally, these proteins would be individually isolated, as glareosin was, to sequence *de novo*. However, due to the low abundance of these proteins, and the dominance of glareosin in male bank vole samples, this was not performed due to constraints of sample volume and of time. Instead, a global proteomics approach was taken to increase the depth of peptide data gained.

3.4.7 Global proteomics of bank vole urine

Protein contentof bank vole urine from 10 different males (n=10) and 10 different females (n=11) was normalised to creatinine (1 µg) to correct for urine dilution and digested in-solution with trypsin. Resulting peptides (between 1.5 – 38 ng, or 0.09 – 2.20 pmol assuming glareosin is the predominant protein component) were analysed by LC-MS/MS and analysed in PEAKS™.

Peptide data were searched against a database of all proteins from *Rodentia* species in the SwissProt database (Bateman *et al.*, 2017) to identify proteins, as for in-gel data, and label-free quantification was performed on the consequent identifications. However, the use of a database containing proteomes of multiple species had some drawbacks for label-free quantification. This often results in multiple identifications made for the same protein, but from different species, which cannot always be tackled with protein grouping (Table 3.1). This presented an obstacle for label-free quantification, as multiple identifications of the same protein from different species that has not been grouped correctly (for example, is identified by some shared peptides but also some peptides from different regions of the protein that are unique to each species) reduces the number of unique peptides per protein and therefore the number of quantifiable peptides. To assess if quantification could be improved by searching against a single-species database, data were also searched against a database of sequences in SwissProt (Bateman *et al.*, 2017) from *M. musculus* that incorporated the three *M. glareolus* OBP sequences, glareosin, and trypsin for autolysis products. *M. musculus* was chosen due to the high number of curated protein sequences available in SwissProt, in addition to homology with *M. glareolus*. Proteomes from other cricetid rodents were not as extensive as that for the house mouse.

PEAKS™ SPIDER searches were used for identification, as it allows for mutations which is beneficial for cross-species matching. A 1% FDR was applied to both sets of results, which set a peptide -10lgP score of ≥35.6 and ≥50.4 for the *Rodentia* and *M. musculus* databases, respectively. The AScore threshold set for PTMs was 20, predicted mutations were accepted with an ion intensity threshold of ≥5%, and a protein score of ≥20 was set, with 1 or more unique peptides. *De novo* only peptides were accepted with an ALC (%) score of 50% or more. A number of parameters were used to consider the success of identification from each database, including the number of peptide spectral matches (PSMs) (the total number of MS/MS spectra matched to a peptide from any identified protein), total number of peptides identified, number of proteins identified and the number of protein groups (proteins identified from the same group of peptides).

Table 3.2 | Summary of protein identifications using cross-species matching against a multiple-species database and a single-species database.
Peptide data generated from in-solution digestion of 10 male and 10 female bank vole urine samples were searched against two databases comprised of protein sequences from SwissProt (Bateman *et al.*, 2017). The first contained all reviewed proteins from the order *Rodentia*, and the second contained only reviewed proteins from the species *Mus musculus*. PEAKS^{TM} SPIDER searches were used to identify proteins.

Database	*Rodentia*	*Mus musculus*
# MS/MS Scans	183889	183889
#PSMs	719	439
#Peptide sequences	258	150
#Proteins	351	44
#Protein groups	89	39
#Protein groups *(manually established, after removal of contaminants)	59	30
De novo spectra	1706	1968

A higher number of PSMs, peptides, proteins and protein groups were established for the *Rodentia* database than for the *M. musculus* only database (Table 3.2). 351 proteins were identified in the *Rodentia* database, compared to 44 in the *M. musculus* proteome, which were grouped into 89 and 39 protein groups by the software, respectively. Of these proteins identified, the *M. musculus* database matched 258 peptides in comparison to 150 from the *Rodentia* database.The *Rodentia* database performed better in terms of the number of identifications possible. However, the protein and peptide scores, sequence coverage, number of unique peptides and number of scans per peptide were considered (Figure 3.17). The distribution of protein and peptide scores were significantly different (Log_{10} transformed data) between *Rodentia* and *M. musculus*, with the *M. musculus* database search scoring higher for both peptide (t-test: $t_{424} = -10.75$, $P \leq 0.001$; difference = 0.14, 95% C.I = 0.12 to 0.17) and proteins (t-test: $t_{393} = -7.1$, $P \leq 0.001$; difference = -0.24, 95% C.I = -0.30 to -0.17). The average -10lgP score for peptides identified from the *Rodentia* database (n = 275) was 58.04 ± 1.25 (mean ± standard error of the mean) and the average score for the *M.musculus* database results (n = 151) was 80.30 ± 2.16 (mean ± standard error of the mean). The average -10lgP score for protein identifications from the *Rodentia* database (n = 351) was 77.43 ± 2.67 (mean ± standard error of the mean) and the average score for the *M.musculus* database results (n = 44) was 137.03 ± 13.28 (mean ± standard error of the mean).The protein coverage was not significantly higher when searching the *M. musculus* database (t-test: $t_{393} = -0.02$, $P = 0.98$; difference <0.01, 95% C.I = -0.13 to 0.13), although the number of unique peptides was significantly different (Log_{10} transformed data; Mann-Whitney U test: U = 10603.500: $n_1 = 44$, $n_2 = 351$, $P \leq 0.001$; difference = 0.00, 95% C.I = -0.301 to 0.000).

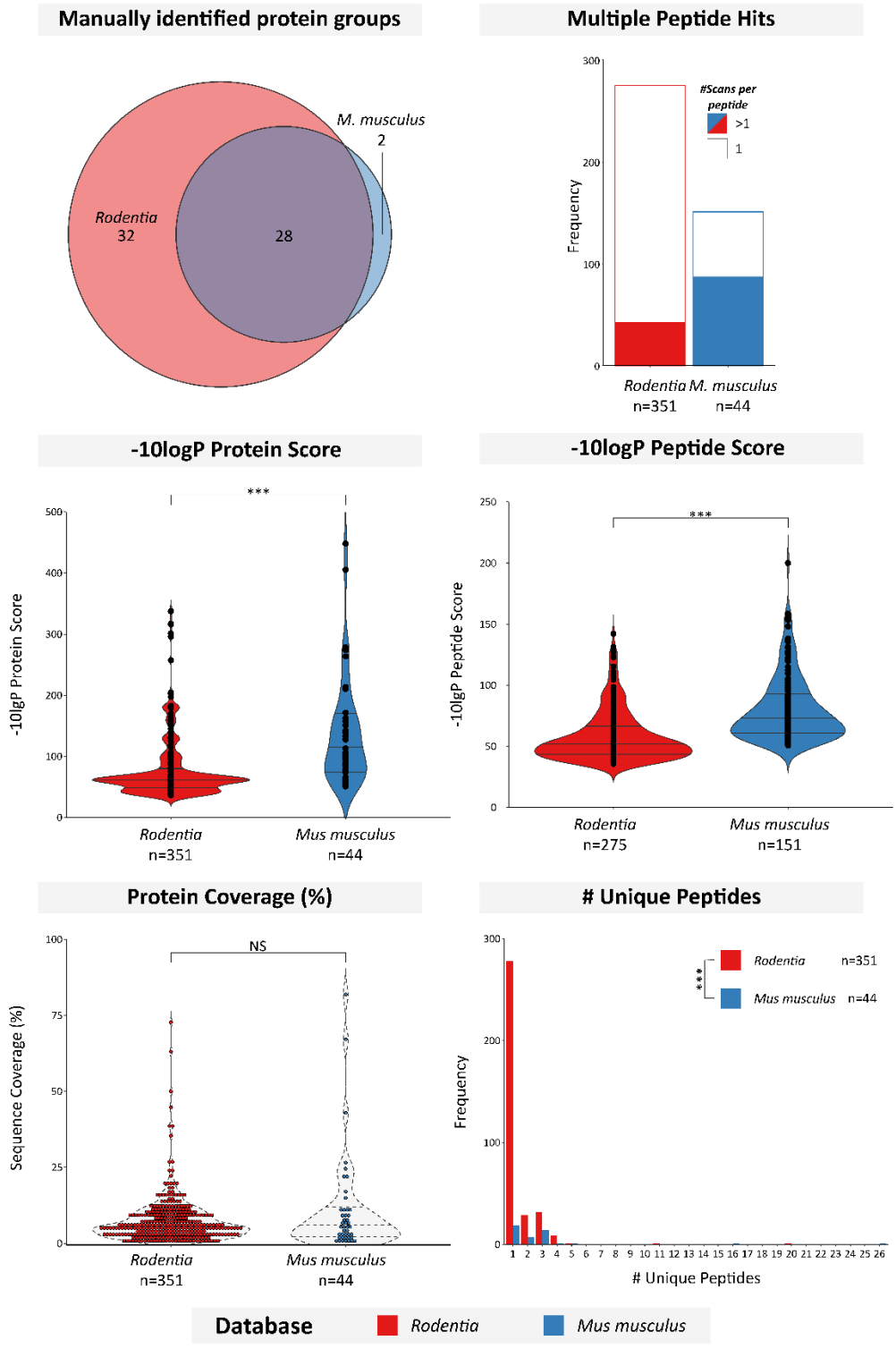


Figure 3.17 | Summary of protein identifications using cross-species matching against a multiple-species database and a single-species database.
Peptide data generated from in-solution digestion of 10 male and 10 female bank vole urine samples were searched against two databases comprised of protein sequences from SwissProt (Bateman *et al.*, 2017). The first contained all reviewed proteins from the order *Rodentia*, and the second contained only reviewed proteins from the species *Mus musculus*. PEAKS™ SPIDER searches were used to identify proteins. The protein and peptide scores were compared, in addition to the protein coverage, number of scans per peptide, and number of unique peptides per protein.

Whilst searching against a database of all *Rodentia* sequences in SwissProt resulted in a greater number of identified proteins and peptides, searching against a single organism proteome improved the quality of both the protein and peptide scores, the average number of unique peptides per protein, and the overall number of scans attributed to each peptide, with no statistically significant change to the sequence coverage.

The *Rodentia* database identified 32 more manually curated protein groups, after filtering out trypsin and human contaminant products, than the *M.musculus* database, which unsurprisingly only identified two proteins not found from the *Rodentia* database, likely filtered out through application of a FDR. A total of 28 manually curated protein groups were found from both databases. Consequently, to investigate the types of proteins within bank vole urine, results from searches against the *Rodentia* database were used, and for label-free quantification, results from a search against the single species *Mus musculus* database was used.

Protein identifications from PEAKS^{TM} SPIDER searches were manually refined so that the same proteins identified from different species were grouped together. Tryptic autolysis products and human contamination products (keratin, actin & collagen) were removed from the analysis. A total of 59 protein groups remained, 32 of which were only identified in males, 4 were only identified in females and 23 were identified in both. Of the protein groups identified in both sexes, six of these were identified in three or more male and three or more female samples. Uromodulin, albumin and serotransferrin were already confidently established in both sexes from in-gel data. Ubiquitin was not previously identified, but is a protein expressed in most tissues (Bastian *et al.*, 2008). OBP3 was previously identified from in-gel data of males, and was identified in seven male samples, but four female samples also retrieved peptide hits. OBP1 was not confidently identified from in-gel data, but five female and three male samples from the global proteomics data provided sequence coverage. In contrast, glareosin and OBP2 were almost exclusively observed in male samples, with peptide matching in one female sample each, compared to nine (glareosin) and five (OBP2) male samples, respectively. Furthermore, the 'identification' glareosin and OBP2 in female urine was reliant on a single spectral count each, and is therefore likely to be false. This highlights the trade-off between quality and breadth of identification in cross-species matching. Perhaps an alternative method of analysis could consider the quality of the relevant peptide identification when deciding the minimum number of peptides required for a confident protein identification.

Only four protein groups were exclusively identified in female samples, however all four proteins were only identified from one peptide each. More proteins were identified exclusively in male samples (n=32). Prostatic steroid-binding protein C1, a member of the secretoglobin family and identified with a -10lgP score of 95.83 was observed in four male urine samples. In the rat (UniProt accession P02782), expression is highest in the testis (Bastian *et al.*, 2008), and it forms a major component of prostate gland secretion and is capable in its dimeric form of binding non-polar steroids (Parker, Needham and White, 1982). Another protein exclusively identified in males is prostaglandin-H2 D-isomerase. Despite a low -10lgP score and identification in a single male, it was identified in some in-gel digestions, although was not the top-scoring protein for any specific bands, and a number of unidentified peptides sequenced *de novo* from the in-gel data matched homologous prostaglandin-H2 D-isomerase sequences when searching non-redundant *Rodentia* proteins in the NCBI database using BLAST (Waridel *et al.*, 2007). Although expressed in a variety of tissues, prostaglandin H2-D-isomerase binds hormones, is likely involved in regulation of the male reproductive system and is expressed in the testis, efferent ducts and epididymis and secreted into the seminal fluid (Beuckmann *et al.*, 1999; Samy *et al.*, 2000). Together with the presence of prostatic binding protein this may indicate the origin of these proteins. The remainder of confidently-identified proteins were likely serum proteins involved in immune response and the transport of metal and calcium ions, heme and lipids.

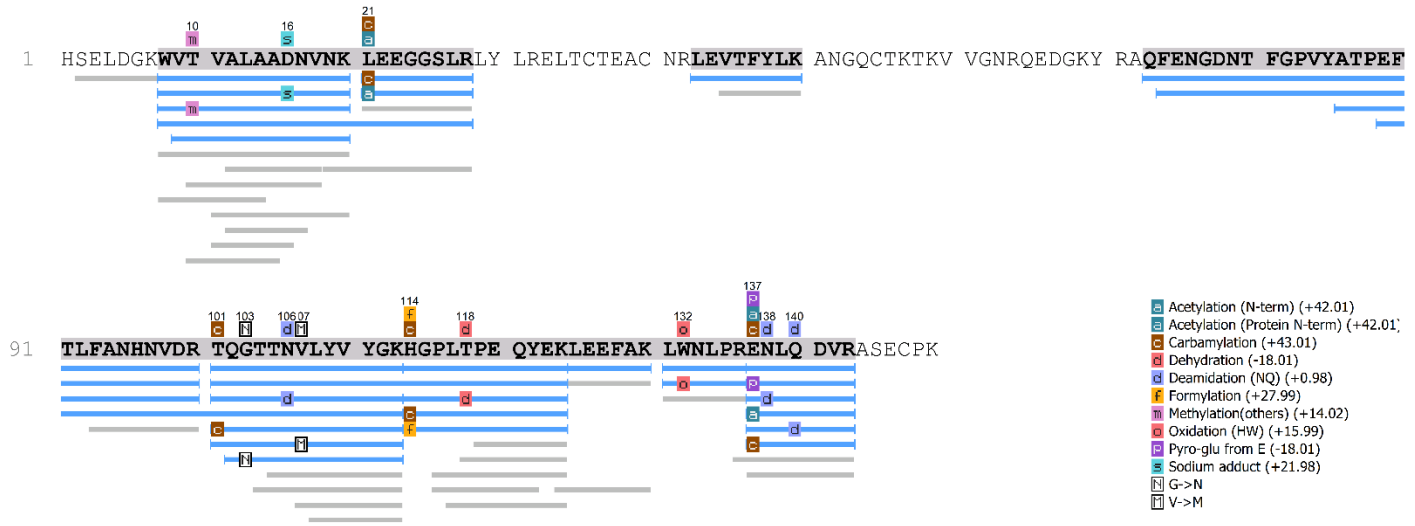
Good sequence coverage for glareosin, OBP3, OBP2 and OBP1 was attained, indicating that all four are likely present in bank vole urine (Figure 3.18). An additional OBP-like protein was also identified from the thirteen-lined ground squirrel, and the same homologous peptides identified from in-gel data were also observed from unidentified *de novo* peptides in the in-solution data.

PEAKS™ Peptide Maps from Global Proteomics Data

Glareosin

-10lgP 261.59

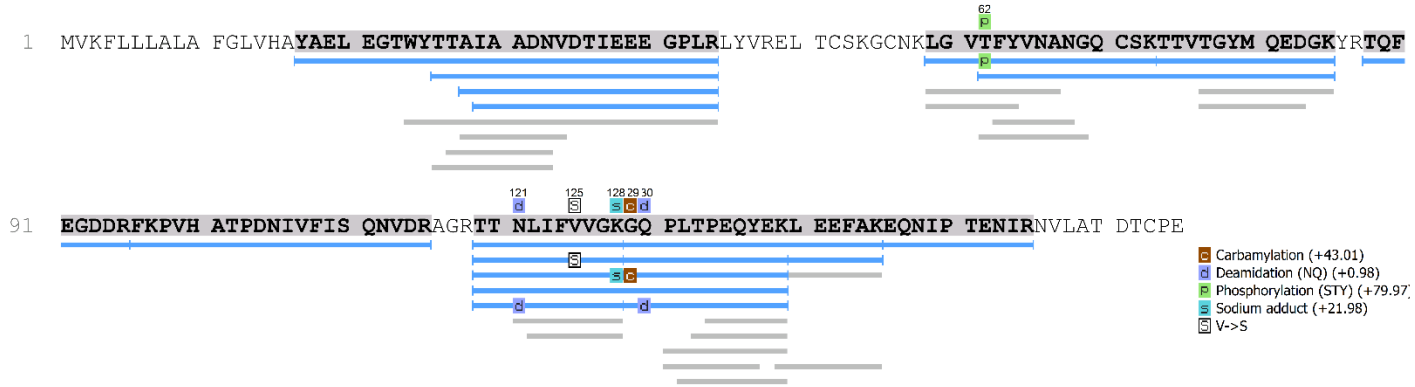
Coverage 67%



OBP3

-10lgP 281.63

Coverage 73%



PEAKS™ Peptide Maps from Global Proteomics Data

OBP2

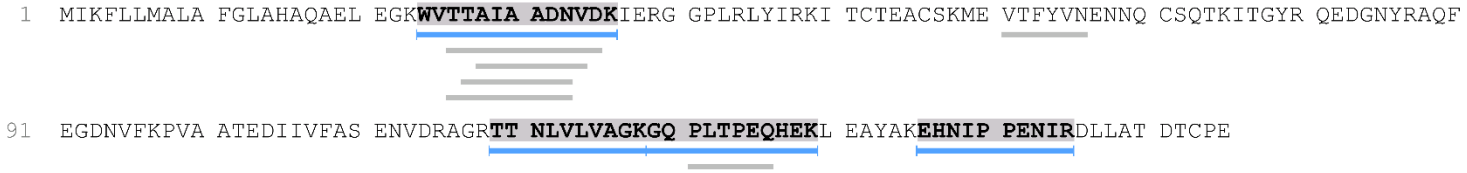
-10lgP 208.40

Coverage 39%

OBP1

-10lgP 177.92

Coverage 27%



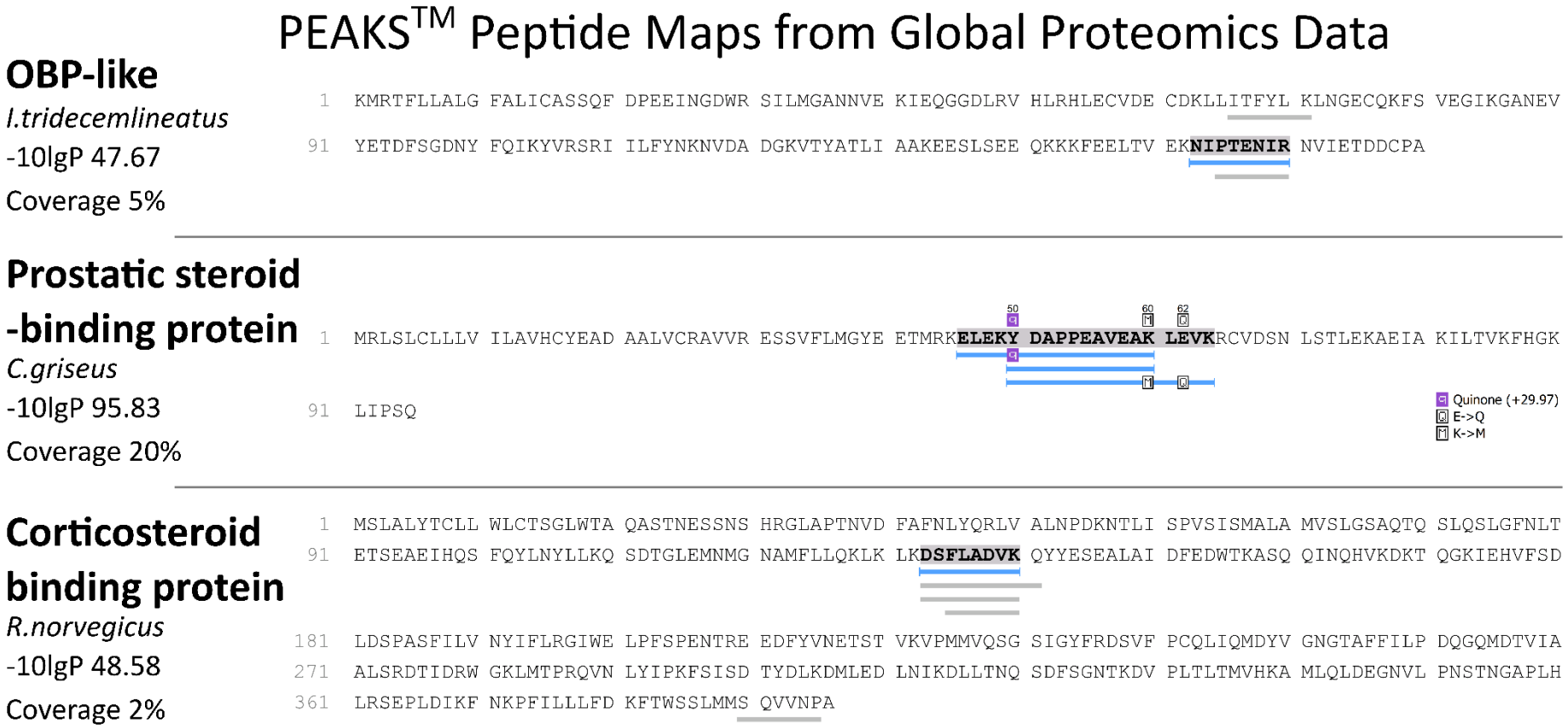


Figure 3.18 | *Sequence coverage of odorant binding proteins.*
Sequence coverage of glareosin, OBP3, OBP2, OBP1 and an OBP-like protein identified from *Ictidomys tridecemlineatus* (thirteen-lined ground squirrel), in addition to prostatic-steroid binding protein (only identified in males) and corticosteroid binding protein (only identified in females) from in-solution digestion and LC-MS/MS analysis of 10 male and 10 female bank vole urine samples.

Label-free quantification results were filtered as recommended by the software vendor. Mass spectra with an associated identification (features) are aligned across all samples to give a feature vector for each peptide. Feature vectors are used to calculate abundance ratios for every peptide across all samples, relative to a chosen sample used as the denominator in all ratio calculations (automatically calculated). It is recommended by the software that feature vectors are filtered either by a quality score, or by peptide area. The quality score is a value calculated by the software that combines m/z error, extracted ion chromatogram shape, retention time error and peptide intensities. The highest quality scores are distributed around an abundance ratio of 1, therefore the software vendor recommends a quality cut-off value that removes most feature vectors with a fold change of more than 8 (Figure 3.19A), therefore adapting the filtering criteria based on the data in question. Another recommended filter is the peptide area, which is filtered in the same way as quality scores; the highest peptide areas are distributed around an abundance ratio of 1, so feature vectors selected for LFQ are taken from those with a peptide area cut-off that selects most vectors with a fold change of less than 8 (Figure 3.19B). No filter was applied to limit the number of samples a protein was required to be identified from, nor was the protein significance or fold change. The number of required unique peptides was set to 1, to account for the limits of cross-species matching, although a peptide ID count of 3 was applied, so that a peptide was required to be identified a minimum of three times.

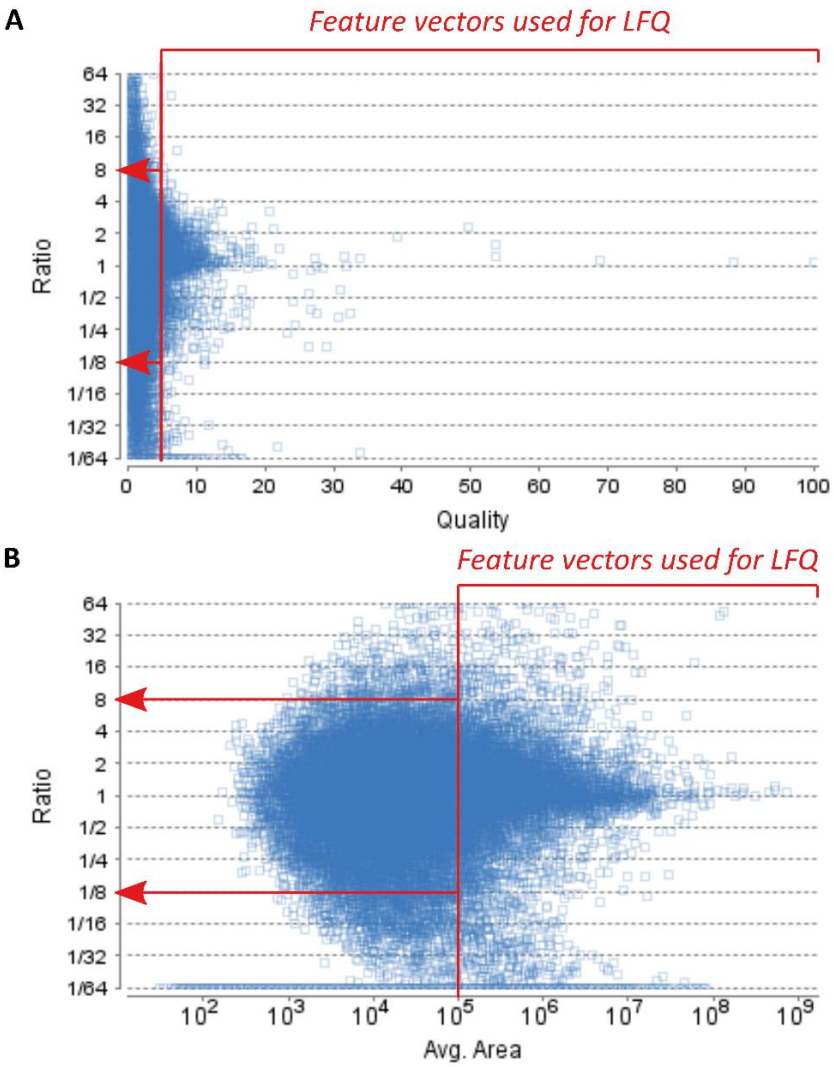


Figure 3.19 | Result-based filtering of feature vectors for label-free quantification in PEAKS™.
The software recommends filtering features, on which label-free quantification will be performed, based on the fold change. Feature vectors with a fold change of less than 8 is recommended, so a quality threshold (A) and feature area (B) is set to reflect this, to select features that are more reproducible and quantifiable. Feature vector graphs taken directly from PEAKS™ label-free quantification results summary and annotated.

After filtering, 294 features were identified, 70 of which were associated with a protein identification. Only a small number of protein groups (n=9) were quantifiable. Three proteins, glareosin, OBP3 and albumin were significantly increased (significance ≥ 25) in males with a fold change of over 2 (Figure 3.20). Uromodulin was the only protein significantly increased in females. OBP1 had on average a higher abundance in male urine, with a fold change just less than two, however the individual values indicate that this is driven by two outlier male samples (Figure 3.21) and the remaining abundance values in male urine appear to be lower than females. Interestingly, these two outlier male samples

appeared more faintly than other male samples when analysed by SDS-PAGE (see Figure 3.2; M2 & M9) and their intact mass profiles were more complex than other male bank vole urine samples (Supplementary). Both profiles indicate unusually high intensity peaks at 16781 and 16802 ± 1 Da.

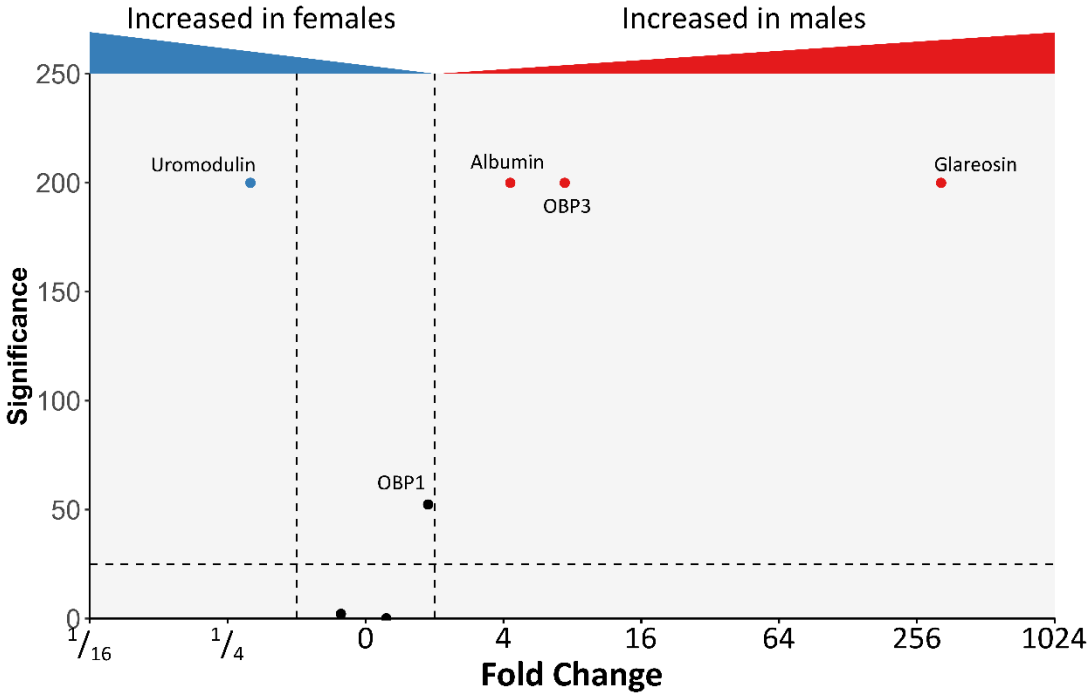


Figure 3.20 | *Label-free quantification.*
Proteins (n=9) in bank vole urine were identified and quantified by PEAKS™. Log_2(Fold Change) is compared with protein significance. Proteins with a significance higher than 25 and a fold change of more than two are indicated.

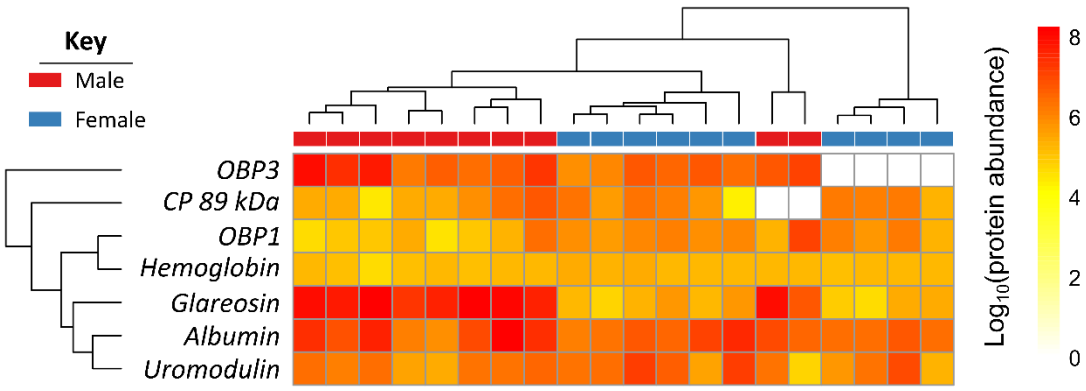


Figure 3.21 | *Label-free quantification.*
Proteins (n=9) in bank vole urine were identified and quantified by PEAKS™. Log_{10}(abundance) was used to cluster both sample and protein. CP 89 kDa, centrosomal protein, 89 kDa.

3.4.8 Preliminary investigation into the protein content of bank vole scent marks

Bank voles use urine deposition in scent mark communication (Johnson, 1975; Christiansen, 1980), but there is some differentiation between 'scent marks' and urine deposits. In captivity, males in particular regularly mark their home territories, and the frequency with which they scent mark affects female mate choice (Kruczek, 1997). Glareosin has been established as the dominating protein in freely expressed urine of male bank voles during the breeding season, with homologous OBPs present at lower abundances (Loxley *et al.*, 2017). However, the protein component of scent marks is unexplored. A preliminary investigation researched the overall complexity and protein expression of scent marks, in comparison to recovered urine, in addition to bladder urine and preputial gland secretions removed during dissection post mortem.

3.4.8.1 Protein content and complexity of bank vole scent marks

Overall protein output and overall complexity were assessed by SDS-PAGE and Bradford protein assay (Figure 3.22). Scent marks (n = 6) were collected from six male bank voles, and after sacrifice, preputial gland secretion (n = 6) and bladder urine (n = 3) were obtained during dissection. Proteins from these samples were compared with those from recovered urine samples from the same males. Scent marks were recovered from a further two females (n = 2) and eleven males (n = 12). The protein concentration of the three bladder urine samples, two female scent mark samples and nine male scent mark samples were measured by Bradford assay, although sample numbers were too low for statistical analysis. Male bladder urine had the highest average protein concentration (2.0 ± 1.0 mg/mL [mean ± standard error of the mean (SE)]). The average protein content in scent marks from males and females was 1.1 ± 0.3 mg/mL and 1.2 ± 0.2 mg/mL (mean ± SE), respectively.

Proteins were separated by SDS-PAGE (volume-normalised, 7.5 µL). Preputial gland (PP) secretions displayed the highest complexity. The three bladder urine (BU) samples varied in overall protein abundance, but four bands resolving at approximately 21, 16, 15 and <6 kDa were consistently strong bands. Recovered scent marks (SM) were generally less complex than the bladder urine and preputial gland samples, and separated to give a similar profile to urine; a number of proteins resolved in a higher molecular weight range (>45 kDa), but the strongest bands were observed between 14 and 21 kDa.

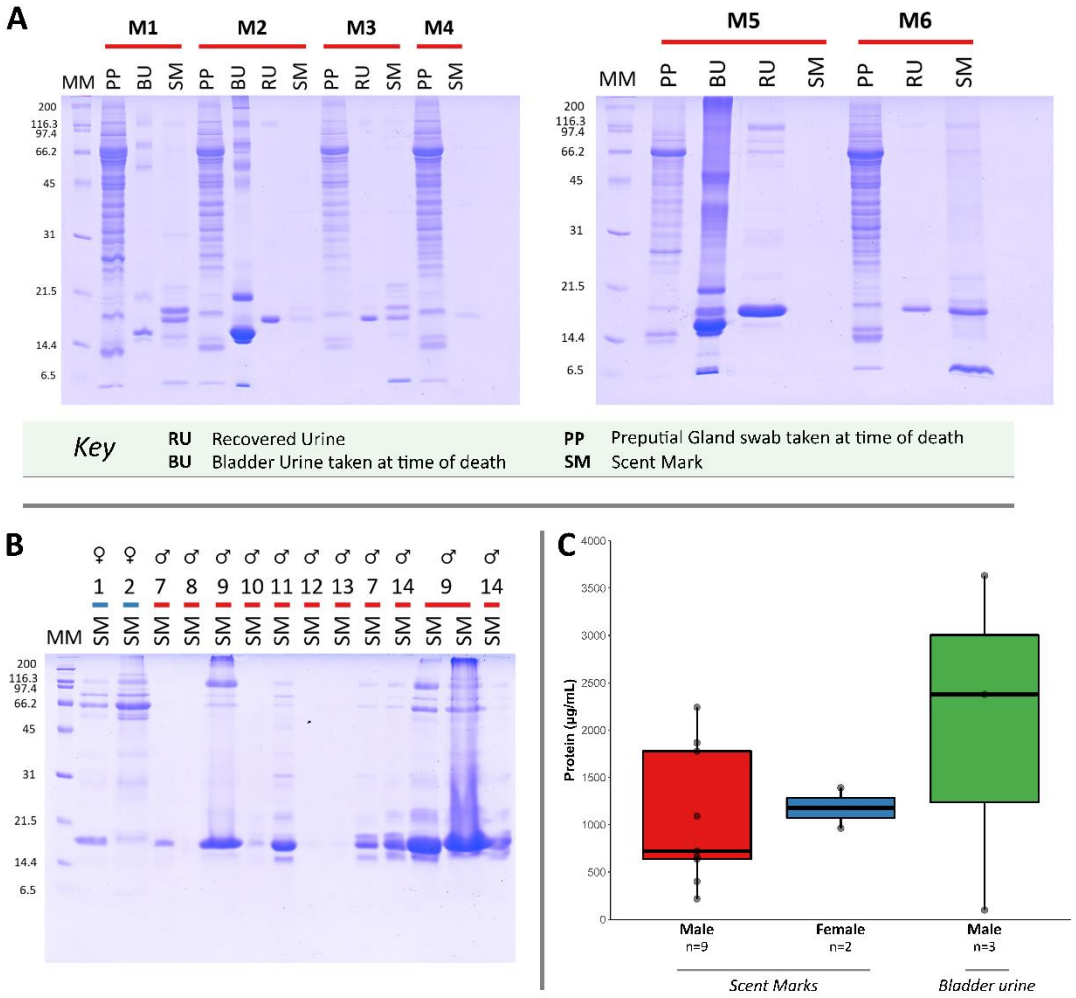


Figure 3.22 | SDS-PAGE and Bradford assay of the protein component of bank vole scent marks.
Scent marks (n = 6) were collected from six male bank voles, and after sacrifice, preputial gland secretion (n = 6) and bladder urine (n = 3) were obtained during dissection. Proteins from these samples were compared with recovered urine samples from the same males (A). Scent marks were recovered from a further two females (n = 2) and eleven males (n = 12) (B). Proteins from each sample were analysed by SDS-PAGE (7.5 µL). The protein concentration of the three bladder urine samples, two female scent mark samples and nine male scent mark samples were measured by Bradford assay (C). Sample numbers were too low for statistical analysis.

The three bladder urine samples, two female scent mark samples and four male scent marks were analysed by intact mass analysis to further investigate protein complexity.

Bladder urine samples were dominated by a single mass, 4871 ± 1 Da (Figure 3.23; (Supplementary), although lower intensity peaks were also observed between 11500 and 13000 Da (11775 ± 1 Da; 12458 ± 1 Da 12498 ± 1 Da and 12969 ± 1 Da).

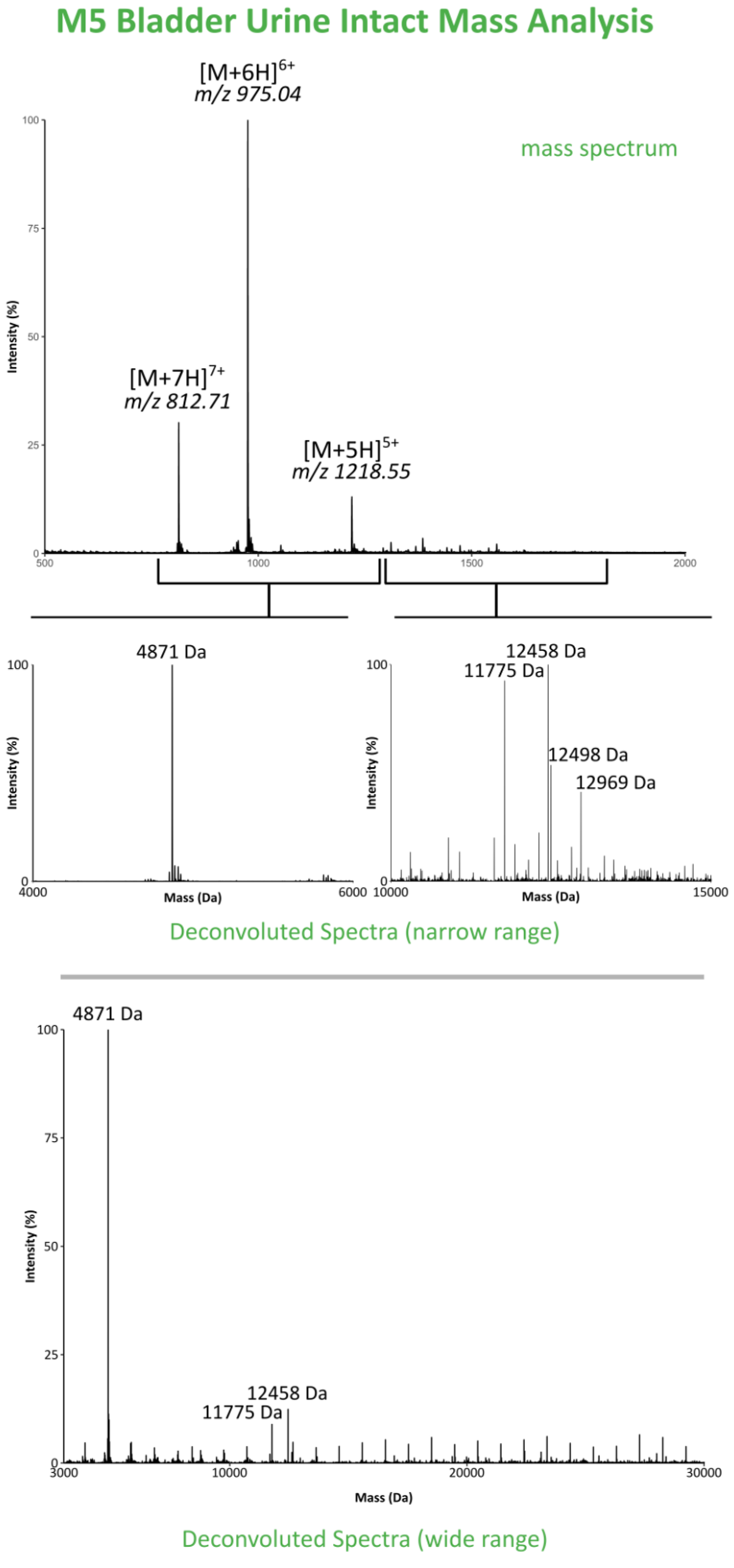


Figure 3.23 | *Intact mass analysis of bladder urine*.
Bladder urine samples (1 pmol) were analysed by ESI-MS to assess overall protein profiles. Deconvolution of the three most abundant peaks in the mass spectrum revealed a single predominant mass, 4871 ± 1 Da. Deconvolution of a higher m/z range also revealed additional peaks (11775 ± 1 Da; 12458 ± 1 Da 12498 ± 1 Da and 12969 ± 1 Da) in all three samples.

Protein profiles of scent marks of both female (n = 2) and male (n = 3) bank voles were also generated (Figure 3.24). Both female profiles contained a predominant peak at 16624 ± 1 Da, whereas all three male samples analysed contained a protein peak of 16764 ± 1 Da, which was the highest intensity peak in two samples. The third sample had additional peaks of 16722, 16929 (glareosin) and 15868 ± 1 Da.

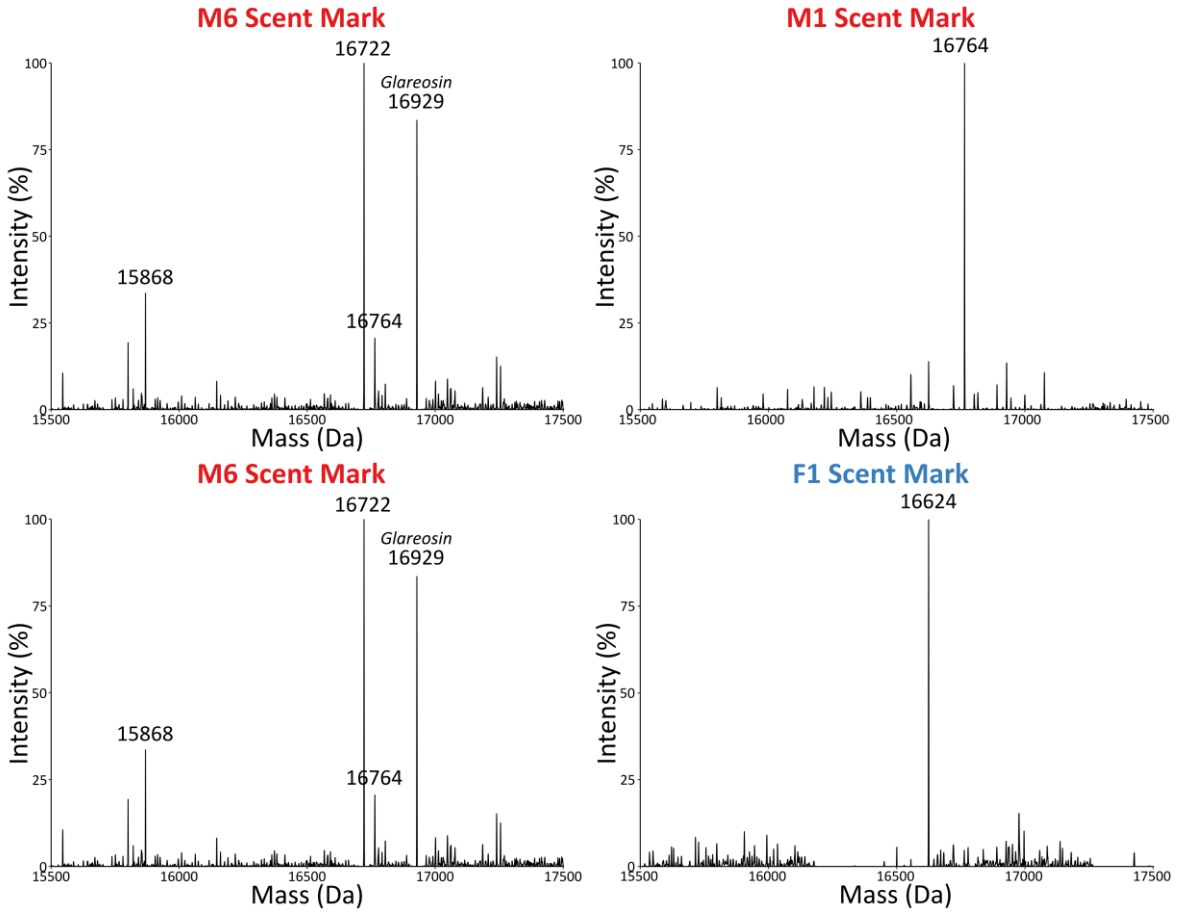


Figure 3.24 | Intact mass analysis of bank vole scent marks.
Scent mark proteins (1 pmol) from 2 females (1 shown, other in supplementary) and 3 males were analysed by ESI-MS. Female scent mark protein profiles were both dominated by a single peak, 16624 ± 1 Da. The protein profiles of male scent marks all contained a peak at 16764 ± 1 Da, which dominated two spectra (M1;M14). One male scent mark also contained the peaks 16722, 16929 (glareosin) and 15868 ± 1 Da. Intact mass analysis of additional samples was not possible.

Protein profiling of bank vole scent marks and bladder urine reveals a low level of protein complexity. Bladder urine samples were dominated by a single peak of 4871 Da, female scent mark samples displayed only one peak of 16624 Da, and male samples consistently contained a mass of 16722 Da, with a more heterogeneous profile in only one sample. It was not possible to analyse further samples due to contamination of poly(ethylene) glycol that dominated the intact mass spectra, thought to be from the sampling method (cotton swabs of cage-deposited scent marks), and a larger-scale analysis would require method optimisation.

3.4.9 In-gel proteome analysis of bank vole scent marks

Intact mass analysis of bladder urine and scent mark proteins generated profiles that were dominated in most cases by a single mass. However, SDS-PAGE analysis suggested a more complex profile. To identify gel bands, proteins separated by SDS-PAGE were excised, and subject to in-gel digestion with trypsin, prior to LC-MS/MS analysis. Protein bands resolving at molecular weights above 45 kDa were not analysed due to the complexity of the samples, and the decision was made to focus on a molecular weight range likely to contain proteins of a similar size to proteins with semiochemical functions, such as MUPs and OBPs. Resulting peptide data were analysed in PEAKS^{TM} and searched against a database of all *Rodentia* protein sequences in the SwissProt database. Top-scoring proteins were considered candidates for identification, which were manually assessed by PEAKS^{TM}-generated *de novo* tags. *De novo* tags are sequences generated *de novo* from experiment spectra by the software matching a given number of amino acids (this work, 5) from a database peptide, without providing a full match. They are therefore a useful consideration when performing cross-species matching as they can suggest homologous peptides.

Analysis of three protein bands each (n = 6) from SDS-PAGE of female scent marks presented peptide-level evidence for vesicular integral membrane protein (resolving at approximately 35 kDa), prostaglandin H2 D-isomerase (resolving at approximately 30 kDa) and all four glareosin/OBP sequences (from a single protein band resolving at approximately 18 kDa) that could not be discriminated between (Figure 3.25). Vesicular integral membrane protein is a calcium-binding lectin capable of binding oligo mannose-type glycoproteins in the membrane of kidney cells (Hara-Kuge *et al.*, 2002; Satoh *et al.*, 2007). In mice, prostaglandin-H2 D-isomerase catalyses the conversion of prostaglandin H2

to prostaglandin D2, and along with several functions in the central nervous system (Hoffmann *et al.*, 1996; Taniike *et al.*, 2002), is likely involved in the maturation and development of sperm in Sertoli cells of the testis (Gerena *et al.*, 2000; Samy *et al.*, 2000).

Analysis of the protein band resolving at approximately 18 kDa presented peptide evidence for all three OBP sequences (Stopková *et al.*, 2010a) and glareosin. OBP1 was the highest-scoring in both gel bands excised. The presence of these proteins contrasts dramatically with the protein profile for female urine. Whereas female bank vole urine has a uniformly low protein expression, female scent marks not only exhibit a predominant protein of 16624 ± 1 Da from intact mass analysis, but in-gel data present evidence for both OBP sequences and glareosin, the expression of which in urine is strongly sexually dimorphic. It is also the only instance that OBP1 is the highest-scoring from in-gel analysis of bank vole urine and scent marks, and the mass observed from ESI-MS, 16624 ± 1 Da, could be a mutated version of the Stopková et al. (2010a) OBP1 sequence (16643 Da).

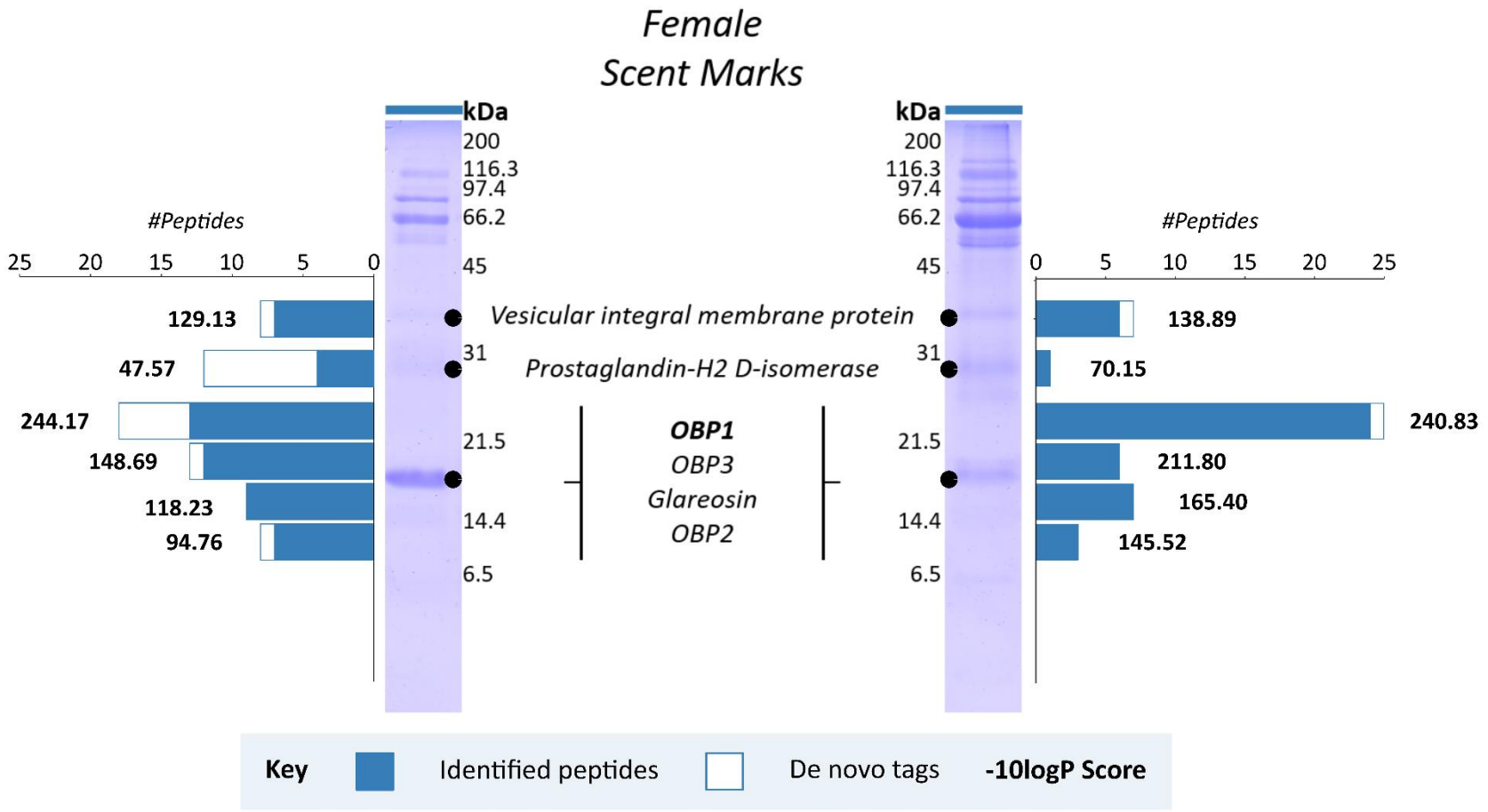


Figure 3.25 | Identification of proteins in female bank vole scent marks by in-gel digestion and LC-MS/MS analysis.
Proteins in two female scent marks were separated by SDS-PAGE. Strong protein bands resolved below 45 kDa were excised, subject to in-gel protease digestion and analysed by LC-MS/MS. Data were analysed by PEAKS^{TM} and identified by searching against a database of *Rodentia* protein sequences from UniProt. Top scoring proteins (-10lgP score, bold) were selected as candidates for potential identification, and PEAKS^{TM}-generated *de novo* tags.

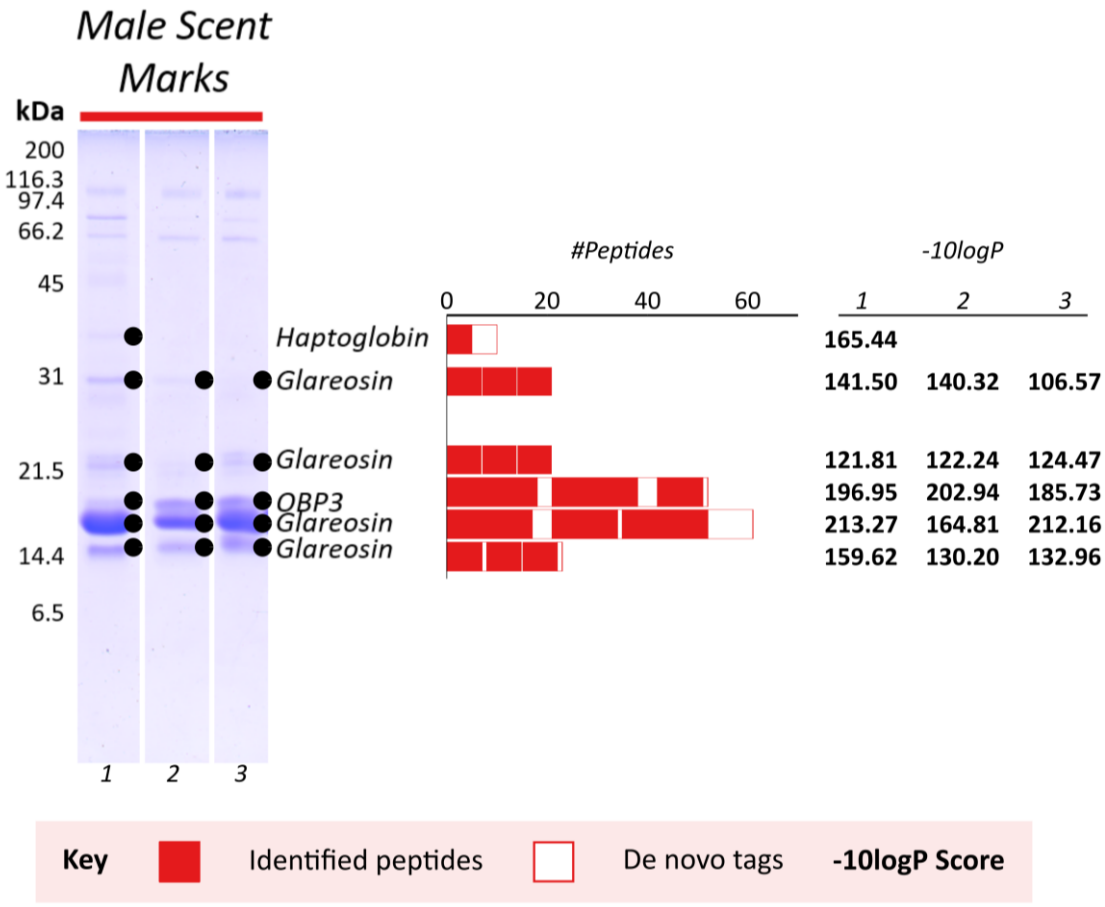


Figure 3.26 | Identification of proteins in male bank vole scent marks by in-gel digestion and LC-MS/MS analysis.
Proteins in male scent marks were separated by SDS-PAGE. Strong protein bands resolved below 45 kDa were excised, subject to in-gel protease digestion and analysed by LC-MS/MS. Data were analysed by PEAKS^{TM} and identified by searching against a database of *Rodentia* protein sequences from UniProt. Top scoring peptides (-10lgP score, bold) were selected as candidates for potential identification. Number of peptides for each sample at a particular molecular weight are stacked (L-R).

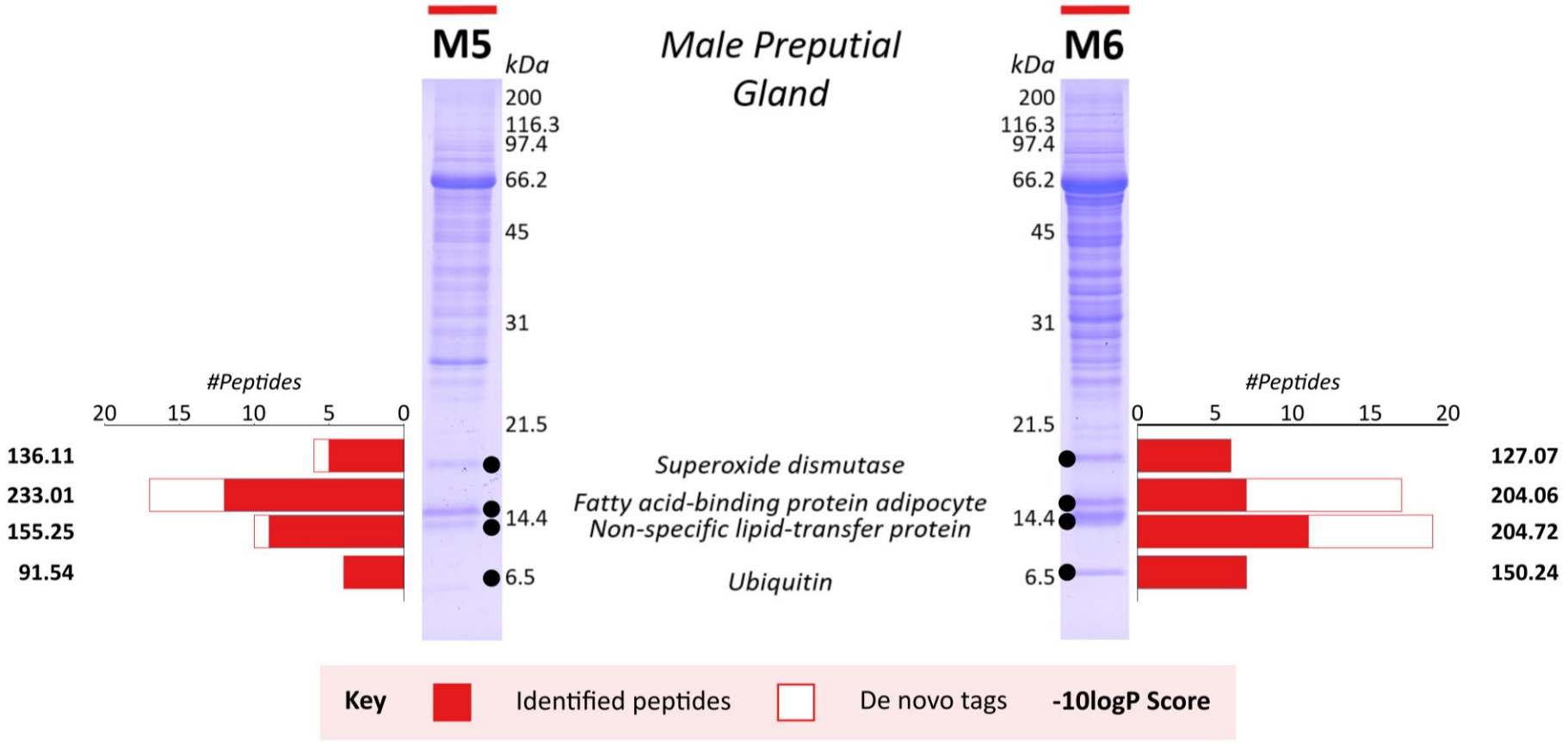


Figure 3.27 | Identification of proteins in male bank vole preputial gland secretion by in-gel digestion and LC-MS/MS analysis.
Proteins in male preputial glands were separated by SDS-PAGE. Strong protein bands resolved below 45 kDa were excised, subject to in-gel protease digestion and analysed by LC-MS/MS. Data were analysed by PEAKS^{TM} and identified by searching against a database of *Rodentia* protein sequences from UniProt. Top scoring peptides (-10lgP score, bold) were selected as candidates for potential identification.

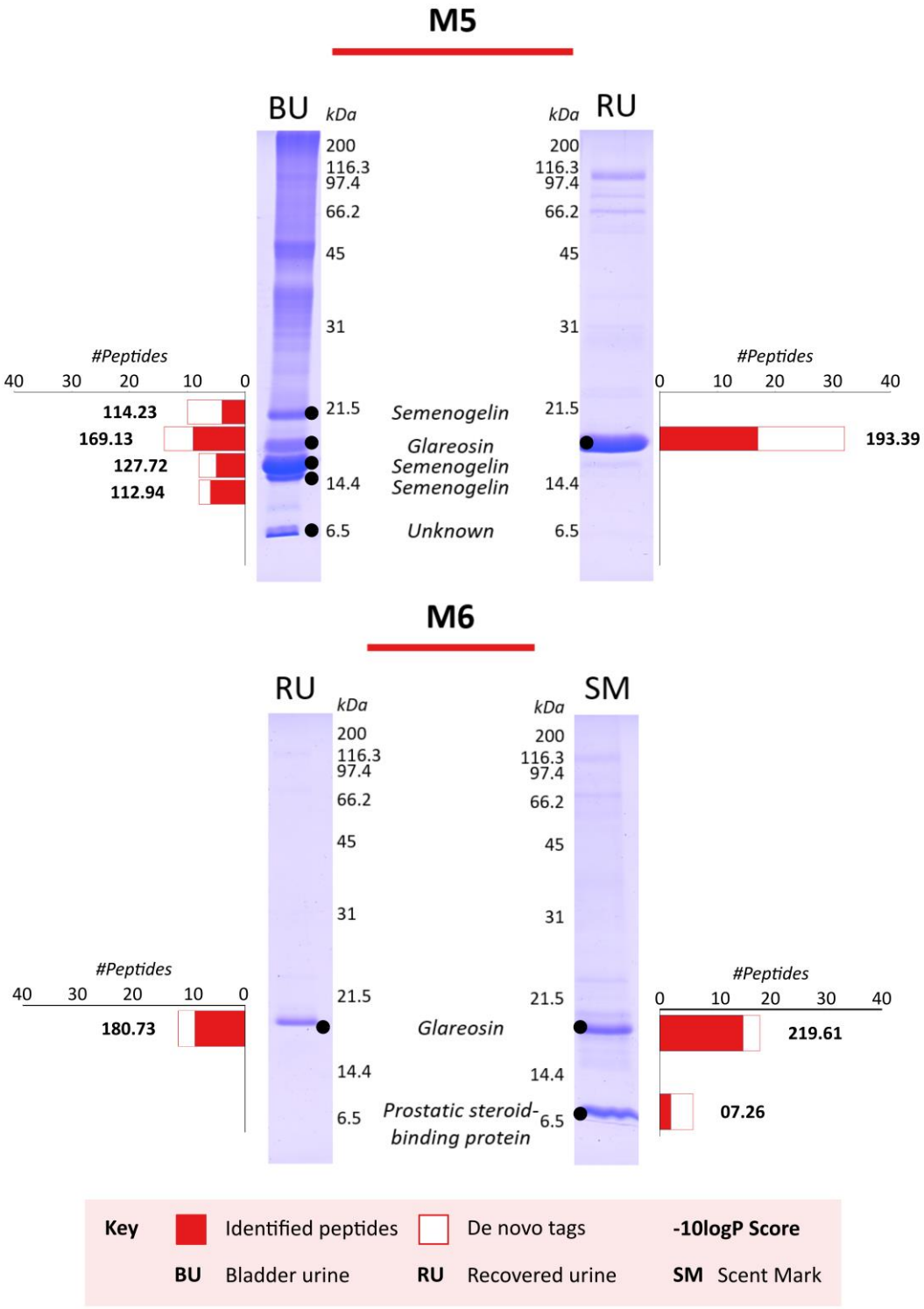


Figure 3.28 | Identification of protein in male bank vole bladder urine, recovered urine and scent marks by in-gel digestion and LC-MS/MS.
The protein content of male bladder urine and a male scent mark were compared to recovered urine from the same males were separated by SDS-PAGE. Strong protein bands resolved below 45 kDa were excised, subject to in-gel protease digestion and analysed by LC-MS/MS. Data were analysed by PEAKSTM and identified by searching against a database of Rodentia protein sequences from UniProt. Top scoring peptides (-10lgP score, bold) were selected as candidates for potential identification.

Proteins in male bank vole scent marks from three bank voles (n = 3) were separated by SDS-PAGE, subject to in-gel tryptic digestion and LC-MS/MS analysis and identified using PEAKS^{TM} (Figure 3.26). A single protein band resolving at approximately 35 kDa was identified as haptoglobin, a plasma protein that binds haemoglobin (Allison and Rees, 1957). The remaining protein bands, resolving at approximately 31 kDa, 22 kDa, 19 kDa, 18 kDa and 15 kDa, were identified as either glareosin (16934 Da) or OBP3 (16753 Da). However, similar to the female scent mark protein band at 18 kDa, these were simply the highest-scoring matches, and peptide level evidence for both OBP2 (16841 Da) and OBP1 (16647 Da) were also observed. It is therefore possible that these bands are a combination of these proteins, but also possible that an identification was not made.

Preputial gland secretions were removed post-mortem and analysed by SDS-PAGE. Protein output was more complex than urine and scent mark proteins, and protein bands below 21 kDa were focussed upon to efficiently compare protein expression in the region with other secretions discussed. Protein bands (n = 8) from two samples were excised and analysed by LC-MS/MS. Fatty acid binding protein (adipocyte) and non-specific lipid-transfer protein were both confidently identified and have roles in lipid transportation. Superoxide dismutase and ubiquitin were also confidently identified, which are both ubiquitously expressed proteins in mammalian tissues (Hershko, Ciechanover and Varshavsky, 2000; Wang *et al.*, 2018).

One bladder urine sample, collected directly from the bladder during dissection, was compared to recovered urine from the same male (Figure 3.27, M5). The single predominant protein band in recovered urine was glareosin, consistent with previous analyses, and the same identification was made for the equivalent resolving protein band in bladder urine. For the fastest-resolving protein band in bladder urine, no identification could be made, but for the remaining three excised bands, peptide evidence was the strongest for semenogelin 1 (also known as seminal vesicle protein secretion 2 (Svs2), or semenoclotin), which strongly suggests seminal vesicle secretion or contamination.

One scent mark was also compared to the recovered urine from the same male. Again, glareosin was confidently identified in the 18 kDa gel band from both samples. A single gel band in the scent mark, resolving at approximately 7 kDa, was identified as prostatic steroid-binding protein, which forms a major component of prostate gland secretions. It is unclear why the 7 kDa protein was not observed in other scent marks, but the SDS-PAGE gel from which most excised protein bands for scent mark analysis were removed from did not

show any protein bands in this region, in comparison to the other two gels (Figure 3.28) and may be a consequence of protein loss from the bottom of the gel when run for too long.

3.4.10 Global proteome analysis of bank vole scent marks and bladder urine

Scent marks from male (n = 5) and female (n = 2) bank voles, and bladder urine samples obtained post-mortem direct from the bladder of two male bank voles (n = 2) were subject to protein-normalised proteolytic digest with trypsin and glu-C (separate digests). The resulting peptides (500 fmol) were analysed by LC-MS/MS and analysed in PEAKS^{TM}. Peptide data were searched against all reviewed *Rodentia* protein sequences in the UniProt database. Protein identifications were not confident enough for label-free quantification, however identifications from PEAKS^{TM} SPIDER searches are discussed.

Few protein identifications were made from any samples analysed, with 22 protein groups identified in bladder urine samples, and 8 protein groups each identified in male and female scent marks (Table 3.3). Of these, only OBP1 was identified in both. Many more spectra were sequenced *de novo* than identified but were not matched.

Table 3.3 | Identification results from LC-MS/MS analysis of bank vole bladder urine and scent marks.
Bank vole bladder urine (n = 2) and scent marks from female (n = 2) and male (n = 5) were subject to in-solution with trypsin and glu-C and analysed by LC-MS/MS. Protein identification was performed in PEAKS^{TM} by searching against a database of *Rodentia* protein sequences from UniProt. The results are summarised below.

Parameter	Bladder urine	Female Scent Marks	Male Scent Marks
MS Scans	21169	20661	50982
MS/MS Scans	8482	9495	25019
Peptide-Spectrum Matches	65	15	31
Peptides	46	14	22
Protein groups	22	8	8
Proteins	83	8	16
De novo sequences generated	445	1294	2691
-10lgP Score at 1% FDR	≥40.8	≥52.2	≥40.8

Analysis of bladder urine samples revealed a number of proteins expressed in cartilage (aggrecan core protein, cartilage oligomeric matrix protein, C-type lectin superfamily member 1 (cartilage-derived), cartilage intermediate layer protein 1 preprotein, biglycan & chondroadherin). The proteins protein-glutamine gamma-glutamyltransferase 4, which plays a role in seminal coagulation (Paonessa *et al.*, 1984), and protein FAM47B, expressed in sperm (Bastian *et al.*, 2008), were also identified, supporting the in-gel peptide data that bladder urine contains seminal secretion.

Glareosin was also identified (Figure 3.29), confirming the identification made from in-gel data. Inspection of unmatched *de novo* sequences by searching sequenced peptides against the NCBI database of *Rodentia* protein sequences revealed a number of highly abundant, high quality sequences homologous to seminal vesicle secretory protein 2, predominantly matching accessions XP_012982014 (*M.auratus*) and ERE71623 (*C.griseus*).

Seminal vesicle proteins 1, 2, 4 and 6 (Svs1, Svs2, Svs4 and Svs6) have all been previously identified by cross-species matching in the seminal fluid of *M. glareolus*, spanning a wide range of molecular weights (approximately 100, 40, 20 and 15 kDa) that are a result of extensive repetitive gene expansion (Ramm *et al.*, 2008) that correlate with the in-gel protein identifications made at the lower molecular weight region from male bladder urine. Repetitive expansion of the central gene region of Svs2 is involved in cross-linking and consequent formation of the copulatory plug (Ramm *et al.*, 2008). In conjunction with the identifications of protein-glutamine gamma-glutamyltransferase 4 and FAM47B, the homologous identified peptides are strong evidence that these bladder urine samples contain seminal fluid.

Male Bladder Urine: Glareosin peptide coverage

-10lgP = 149.83

Sequence coverage = 25%

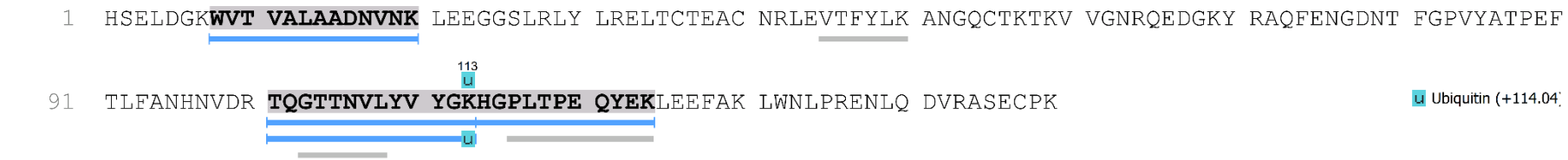


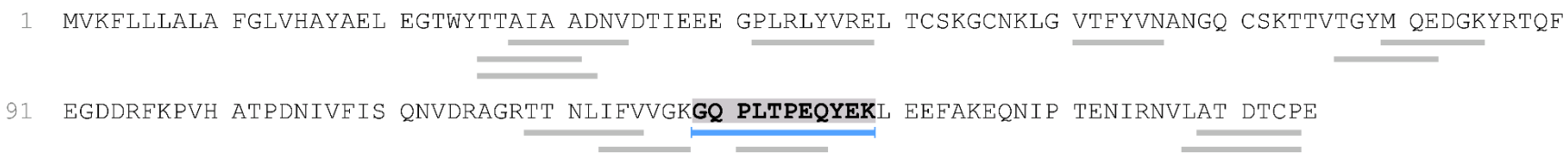
Figure 3.29 | Sequence coverage of glareosin from peptide data of male bladder urine samples.
Male bladder urine samples (n = 2) were obtained during dissection. Proteins were digested in-solution with both trypsin and glu-C in separate digests, and the peptide mixture was analysed by LC-MS/MS. Data were analysed in PEAKS™, where identifications were made by searching against a database of *Rodentia* sequences from UniProt. Glareosin was identified with a sequence coverage of 25%.

Female scent marks were analysed in the same way. Identified proteins (n = 8) included OBP1 and OBP3 (Figure 3.30), in addition to typical urinary proteins albumin, serotransferrin and transthyretin. Cartilage intermediate layer protein (UniProt Accession: I3N356), E3 ubiquitin-protein ligase MYCBP2 (UniProt Accession: G3HI28) and PCF11, cleavage and polyadenylation factor subunit (UniProt Accession: H0V2E3) were the remaining proteins identified. Only albumin and PCF11, which in humans is expressed in both male and female reproductive tissues (Bastian *et al.*, 2008), were identified with more than one unique peptide.

To investigate further, the 1294 unmatched spectra sequenced *de* novo were explored manually. Many of the peptides were identified as *de novo* tags (partially matched to database peptides) for OBP1 and OBP3 (Figure 3.30, grey bars), and these were summarised by comparing the de novo sequences to the established sequences of glareosin and OBPs in addition to the database-matched sequence coverage (Figure 3.31). The *de novo* sequences, although variable in some lower-quality regions which would require manual assessment, give good sequence coverage, identifying observed mutations in the OBP and glareosin sequences. For example, point mutations TTNJJFV[VG→GR]K and TTNJVJ[VA→AR]GK were observed in the experimental data for the peptide sequences TTNLIFVVGK (OBP3) and equivalent TTNLVLVAGK (OBP1) (where 'J' indicates a leucine or isoleucine).

Female Scent Marks: OBP3 peptide coverage

-10lgP = 81.15 *Sequence coverage = 7%*



Female Scent Marks: OBP1 peptide coverage

-10lgP = 76.67 *Sequence coverage = 8%*

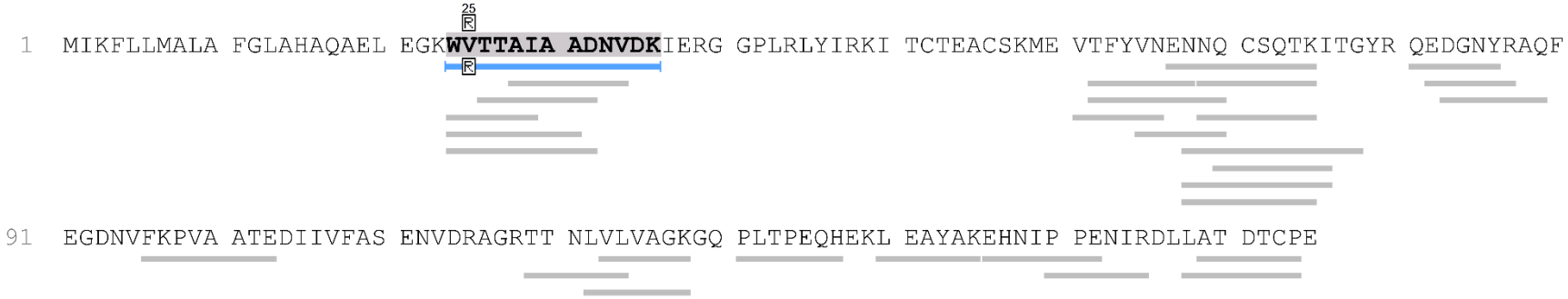
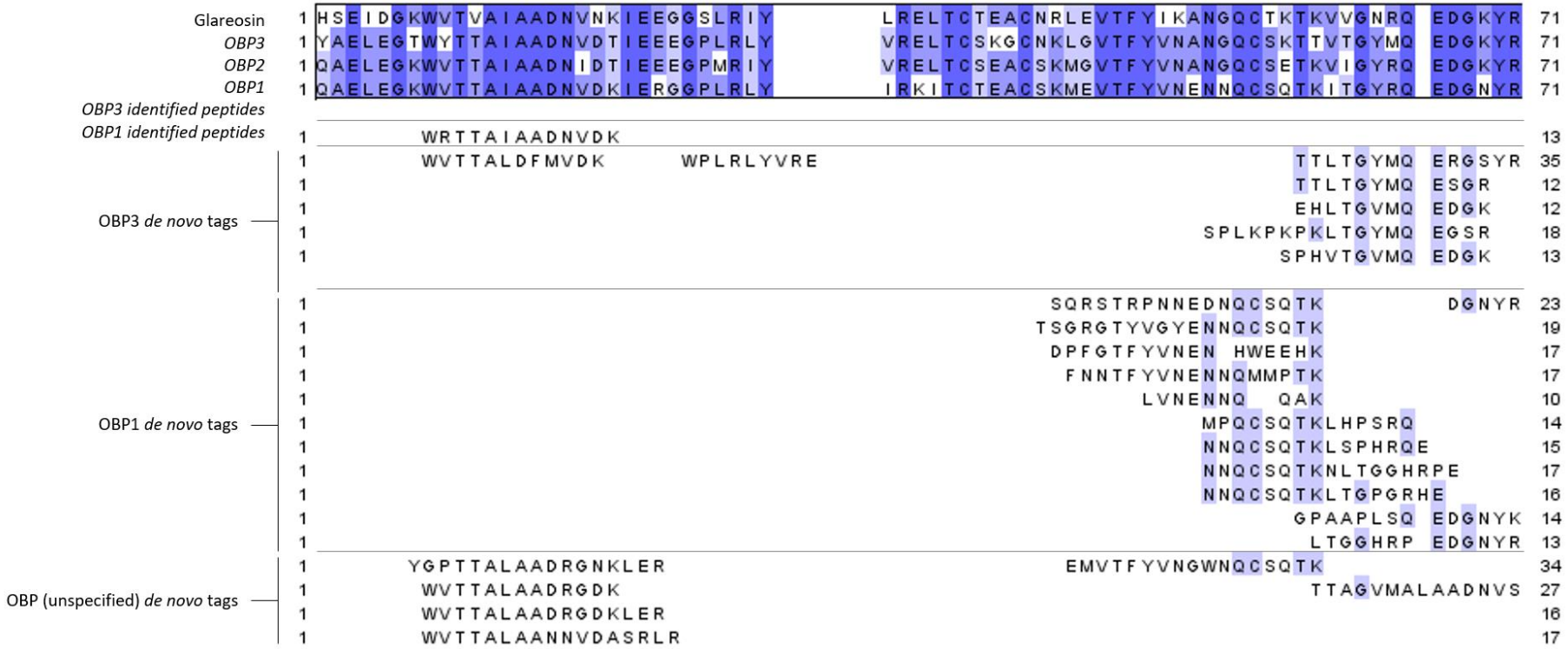


Figure 3.30 | OBP sequence coverage from in-solution digestion and LC-MS/MS analysis of female scent mark samples.

OBP-like Sequence Coverage from Female Scent Mark Peptide Data



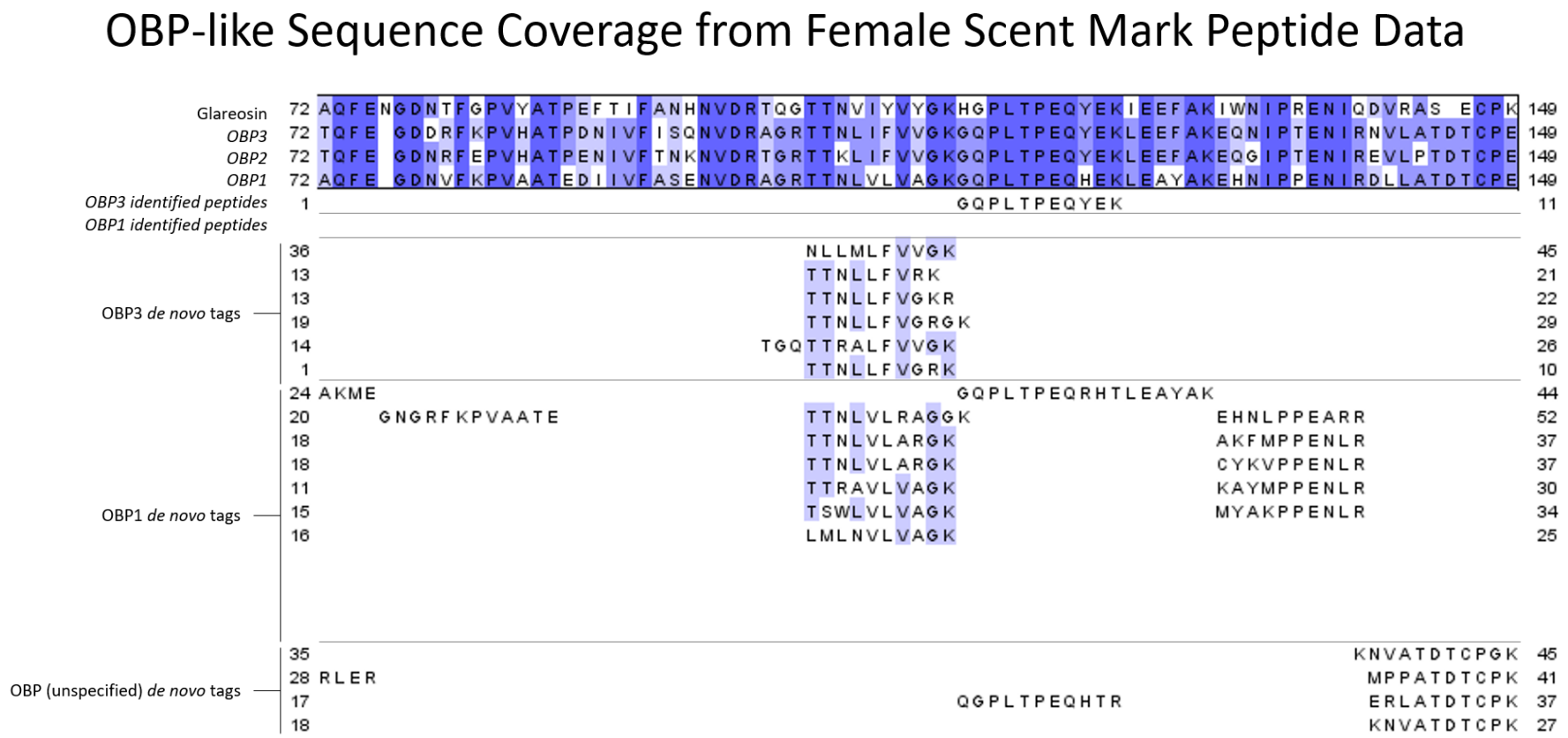


Figure 3.31 | OBP sequence coverage from de novo tag spectra generated from LC-MS/MS analysis of peptides retrieved from in-solution digestion of female scent marks.

Identification of OBP3 and OBP1 from female scent mark peptide data were supported by identified *de novo* tags. Supporting peptides were aligned with the protein sequences of glareosin, and OBP1-3, in addition to the database-matched peptides ('identified peptides'. The alignment was visualised in JalView (Waterhouse *et al.*, 2009).

Eight protein groups were identified from analysis of male scent marks (n = 5), including glareosin, OBP3, OBP1, and a protein group consisting of aphrodisin-like protein (UniProt Accession: A0A061HV58) from *C. griseus* and OBP-like protein (UniProt Accession: S5ZYD3) from *P. sungoras* (Figure 3.32). Other proteins identified were nesprin-1 (UniProt Accession: H0VC35), aggrecan core protein (UniProt Accession: P07897), interleukin enhancer binding factor (UniProt Accession: Q9Z1X4) and protein ZNF365 (group of 8 proteins).

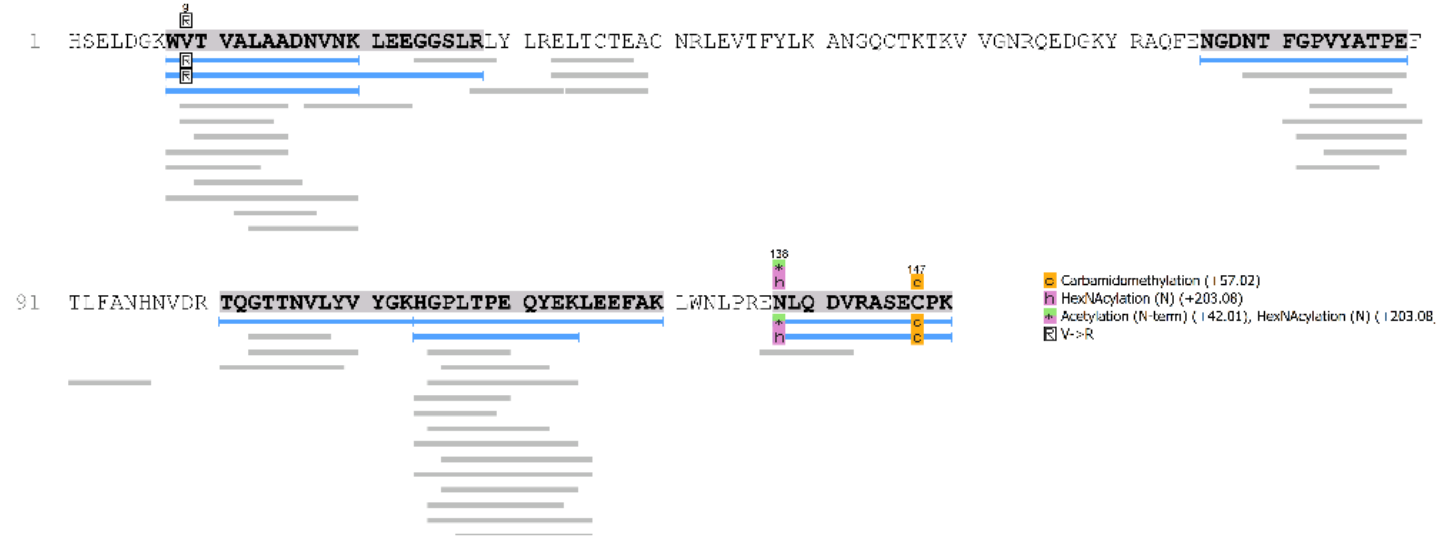
Due to low numbers of proteins identified, the unmatched *de novo* sequences were assessed manually. Like the female scent mark proteins, many were homologous to the OBPs. *De novo* tags of these proteins are indicated by the grey sequence coverage in Figure 3.32.

Male Scent Marks

Glareosin

-10lgP = 156.78

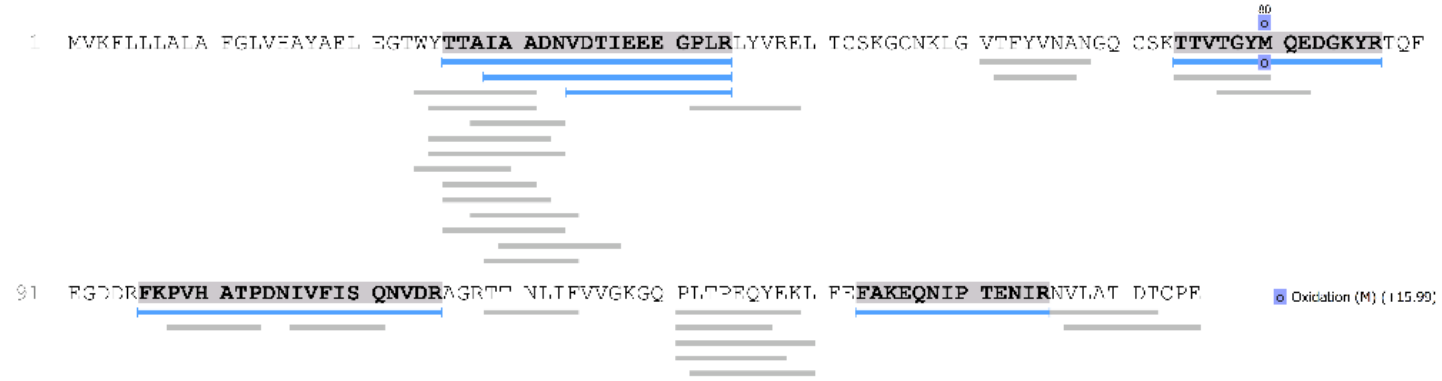
Sequence coverage = 52%



OBP3

-10lgP = 132.80

Sequence coverage = 40%



Male Scent Marks

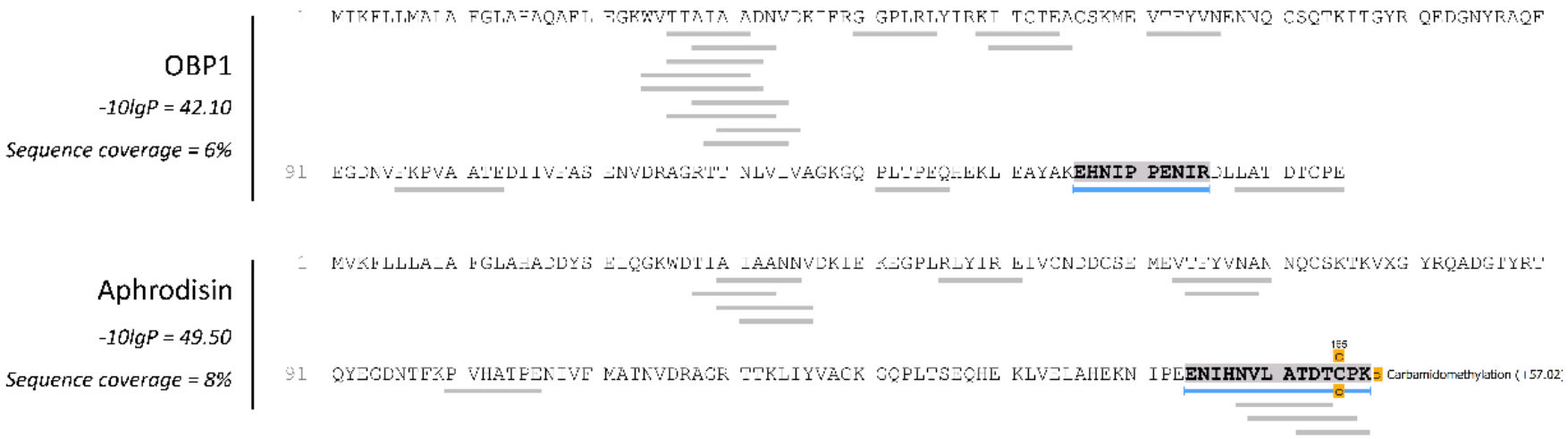


Figure 3.32 | Sequence coverage of glareosin, OBPs and aphrodisin-like protein (C.griseus) from peptide data generated from in-solution digestion of male scent mark samples (n = 5).

Whilst the urinary proteome is dominated by glareosin in breeding season males, and largely absent in females, the scent mark proteome of bank voles is more complex.

Firstly, the protein content of female scent marks has a very different SDS-PAGE profile than their urinary output. SDS-PAGE analysis of scent marks displayed a protein band that resolved at approximately 18 kDa, not seen in urine. A single dominant peak was observed when samples were analysed by intact mass analysis (16624 ± 1 Da). In-gel digestion revealed peptide-level evidence for glareosin and all three OBP sequences, the highest-scoring of which was OBP1, in addition to prostaglandin H2 D-isomerase, which is usually associated with sperm maturation (Gerena *et al.*, 2000). In-solution digestion of scent mark proteins revealed low peptide-level evidence for OBP1 and OBP3, but more importantly generated many abundant, unidentified peptide sequences *de novo* with homology to one or both of these proteins. It is therefore likely that other gene products homologous to these proteins exist. It was not possible to confidently sequence the protein giving rise to the 16624 Da mass peak observed, but it is highly likely this is a new gene product of the OBP1 or OBP3 sequences previously established, either a homologous addition to the pre-established OBP genes, or a mutated OBP sequence synonymous to the previously established bank vole gene product. Purification of the protein, or proteins, would achieve full sequencing, however sample numbers and volume were too low in this analysis.

Male scent marks had approximately the same overall average protein concentration of 1 mg/mL as breeding season urine. SDS-PAGE and in-gel proteomics suggested a similar profile to male urine. Glareosin, OBP3, OBP2 and OBP1 were all identified from in-gel peptide data and it was difficult to discriminate between them for identification of different protein bands. However, intact mass analysis, albeit from a small number of samples, revealed a peak of 16764 Da as the overriding species in two samples, and at a lower level in another sample, which also identified peaks of 15868, 16722 and 16929 (glareosin) ± 1 Da. As for the mass of the female-only protein, these masses did not match those predicted for OBP1, 2 or 3 (Stopková *et al.*, 2010b). In-solution digestion of male scent marks (n = 5) revealed good peptide-level evidence for glareosin, OBP3, OBP2 and an additional OBP-like protein, which shared peptide evidence for both an aphrodisin-like protein from *C.griseus* and an OBP-like protein from *M.auratus*. Many *de novo* tags were identified as homologous to these proteins, and suggest that new gene products of these proteins are highly likely candidates for scent marking proteins.

Bladder urine from male bank voles, collected post-mortem directly from the bladder, was also investigated. Few samples were available, so no solid conclusions were possible, but both samples available to analyse by ESI-MS shared a dominating peak of 4871 Da. One gel lane was available for

in-gel digestion, and glareosin was identified in addition to seminal vesicle protein 2, a seminal protein. This identification was confirmed from in-solution LC-MS/MS analysis, from manually identified peptides selected from high quality *de novo* sequences generated by PEAKS^{TM}, however it is unclear which protein species is responsible for the 4871 Da mass in the protein profiles.

In-solution digestion was not possible to perform on preputial gland secretions, however in-gel digestion did not identify any known lipocalins sequences, either as a direct match or from unmatched manually searched peptides sequenced *de novo*.

3.5 Discussion

Glareosin appears to be the major protein output in male bank vole urine that is stimulated during the breeding season. As a lipocalin with a clearly defined central cavity that could be switchably accessible, combined with male specific production and a seasonal expression pattern, this points to a role for glareosin as a major driver of chemical communication between male and female bank voles. As we gain a better understanding of the use of lipocalins in chemical communication in rodents, an interesting bifurcation is increasingly evident. Of rodents, Muridae (old world mice, rats) have evolved polymorphic families of MUPs that create considerable potential for individual variation in proteins – they function as pheromone binding proteins but also as pheromones in their own right. Currently, our knowledge is largely derived from studies of house mice (*Mus musculus*) and brown rats (*Rattus norvegicus*). By contrast, Cricietinae (hamsters, voles) also elaborate protein in their secretions, but evidence thus far suggests that this is restricted to high levels of a single protein. Thus, roborovskin, from *Phodopus roborovskii*, is a single lipocalin produced in the urine equally by both sexes (Turton *et al.*, 2010). The vaginal discharge of the golden hamster, *Mesocricetus auratus*, contains abundant levels of the lipocalin aphrodisin, which acts as a pheromone (possibly in concert with a bound ligand) to stimulate copulatory behaviour by males (Singer et al. 1986; Vincent et al. 2001; Henzel et al. 1988; Briand et al. 2000; Briand et al. 2004). Aphrodisin is a female-specific lipocalin in vaginal secretions whereas glareosin is a male specific protein restricted to the breeding season. Whilst none of these species invoke the same polymorphic variation as MUPs as the muridae, it is probable that clear functions in intra-specific communication will be found. Interestingly, Bathyerginae (*Fukomys*, naked mole rat) also seem to express urinary proteins that are more aphrodisin-like (Hagemeyer *et al.*, 2011). It is possible that MUP-like sequences have evolved different roles to aphrodisin/OBP like proteins, and that in muroid rodents, a high level of polymorphism may be a unique feature. Whereas MUPs are readily identified and classified within the lipocalin family, there is a need for clearer understanding of the aphrodisin-like proteins. OBPs are expressed in nasal tissue in a wide range of species (Pes and Pelosi, 1995; Garibotti *et al.*, 1997; Briand *et al.*, 2000; Tegoni *et al.*, 2000; Lobel *et al.*, 2001; Heydel *et al.*, 2013) and may facilitate the transport of low molecular weight signalling molecules across the mucosal membrane. However, OBPs are now being increasingly reported in the urine of rodents, and it is likely that they are also involved in the generation as well as the reception of chemosignals. Further study of the role of lipocalins in chemical communication seems likely to reveal a breadth of mechanisms whereby information is conveyed between conspecifics.

Urinary proteome analysis addresses the question of additional protein complexity. Both in-gel and global characterisation of urinary proteins presented peptide-level evidence for OBPs. Substantial

evidence for OBP1 and OBP3 was obtained in both males and females. However, glareosin and OBP2 were only identified by one spectrum each in single, different, female samples, and were only confidently identified in male samples, so it is possible that these two proteins are male-specific, as the PSMs in female samples could be a result of cross-contamination from analysis on the same LC-MS/MS system. However, intact mass analysis did not reveal any masses equivalent to the established OBP sequences. Manual inspection of peptides sequenced *de novo* revealed additional peptides homologous to established OBPs. Homologous peptides covered regions of the protein sequences that were identified by database matching, and were therefore in most cases unlikely to be alternative to the established OBPs, but instead this suggested additional OBP-like proteins were present.

Global proteomics confidently identified prostatic steroid-binding protein (PSBP) in male samples only and corticosteroid binding globulin (CBG) in females only. PSBP is secreted into seminal fluid from the prostate gland (Hurst and Parker, 1983) and may be an indicator that urine contains proteins expressed in the prostate, and therefore suggests a possible origin for other urinary proteins identified. Together with the identification of seminal vesicle protein 2 in male bladder urine, this suggests an overlap between excretion and seminal fluid production, however the discrepancy between protein content of recovered urine and bladder urine suggests this overlap is regulated in some way. CBG is implicated in steroid transport (Lin and Muller, 2010) and is capable of binding glucocorticoids and progesterone (Hammond 1990; Dey & Roychowdhury 2003).

Label-free quantification was performed on eight proteins, and revealed a strong sexual dimorphism in both glareosin and OBP3. Quantification of OBP1 was interestingly driven by two male samples with unusually high abundance ratios of this protein, whereas the abundance in remaining samples tended to be lower than in females. Notably, these two males had intact mass profiles that diverged slightly from the other male samples.

A survey of the protein output of bank vole scent marks revealed relatively simple intact mass profiles. Both female samples and two of three male samples were dominated by a single peak (female, 16624 ± 1 Da; male, 16764 ± 1 Da), although additional peaks were observed in one male scent mark. Likewise, intact mass analysis of male bladder urine was dominated by a single peak, 4871 ± 1 Da. Female scent mark samples, unlike their recovered urine counterpart, had a higher level of overall protein, and presented peptide-level evidence for all three OBPs and glareosin, in addition to prostaglandin-H2 D-isomerase, in contrast to urine, analysis of which only identified peptide-level evidence in more than one sample for OBP1 and OBP3. Furthermore, additional OBP-

like peptides were identified as *de novo* tags in in-gel analyses and global analyses, which suggested another OBP-like protein or proteins.

Male bank vole scent marks resolved on an SDS-PAGE gel in a similar way to urine, and in-gel data identified the same protein profile; high-scoring identifications of glareosin and OBP3, with additional OBP sequences, some of which were novel. Global analysis revealed a similar story, with many novel OBP sequences identified. Notably, none of the novel OBP sequences identified from male urine, female scent marks or male scent marks were identical, suggesting a different novel OBP, or OBPs, in each sample type.

Protein identification and label-free quantification for both urine and scent marks were poor. Whilst global proteomics of urine samples identified a greater number of proteins than from scent marks, many of these were still based on only one unique peptide, or from only one sample. Label-free quantification was therefore only possible on eight proteins. When scent mark data were searched, identification was invariably worse than analysis of urine samples, and label-free quantification was not possible. This is indicative of the issues when performing cross-species matching. If these proteins are involved in scent signalling, and are therefore under much higher evolutionary pressure than housekeeping proteins (such as albumin and uromodulin, which were easily and confidently identified), then these sequences are likely to be more evolutionarily divergent than other proteins when comparing homologous species. Proteomics of these samples, without a sequenced genome, requires a much more careful and manually assessed approach.

3.6 Supplementary

3.6.1 Published paper

3.6.2 Individual intact protein mass spectra

3.6.2.1 Captive bank vole urine

3.6.2.2 Wild bank vole urine

3.6.3 Glareosin sequencing fragment ion data (from paper supplementary)

3.6.4 Glareosin heavy leucine data

3.6.5 Multiple sequence alignment for phylogenetic analysis (from paper supplementary)

3.6.6 Bladder urine intact protein mass spectra

3.6.7 Female scent mark intact protein mass spectra

4 Characterisation of the urinary protein content in the field vole, *Microtus agrestis*

Abstract

The field vole (*Microtus agrestis*) is a rodent of the *Cricetidae* family, common in the northern hemisphere. Whilst closely related to the bank vole, this species exhibits cyclic population changes every three to four years, of which the origin is largely not understood. Whilst the social organisation of these predominantly solitary animals is reported to a limited extent, the communicative mechanisms on which their social structure are based is particularly absent from current literature. Following the investigation into the urinary protein profile of the bank vole, *Myodes glareolus*, of which male members exhibit a sexually dimorphic, seasonally-expressed protein (glareosin), a similar investigative approach was started to survey the complexity of both the urinary and scent mark proteome in male and female individuals of this species. Glareosin-like protein variants were reported in mature male field voles, present in both urine and scent marks, with established amino acid substitutions indicative of a level of heterogeneity higher than that of the bank vole, but lower than that seen in mice and rats. Evidence of a MUP-like protein, potentially glycosylated, was also discovered in both males and females, in mature adults in addition to juveniles. Another lipocalin variant was also discovered related to the bank vole OBP3, and field vole scent marks presented evidence of a further type of lipocalin. Overall, investigation into field vole urine indicates a complex profile, with not only sexually dimorphic, polymorphic expression of glareosin, but also evidence of other lipocalin types in both sexes from both urine and scent marks.

Contributions

Urine and scent mark samples were collected by Holly Coombes (Mammalian Behaviour and Evolution Group, University of Liverpool), who also provided SDS-PAGE gels of field vole samples for use in figures where indicated, and for in-gel digestion for identification. An additional set of urine samples for protein and creatinine assay data wascollected by Alice Maher, also from the Mammalian Behaviour and Evolution Group, University of Liverpool.

4.1 Introduction

The Eurasian field vole, *Microtus agrestis* (Linnaeus, 1761), is a herbivorous rodent that together with lemmings form the subfamily *Arvicolinae*, within the *Cricetidae* family of rodents. The species occupies most of Europe in addition to western Asia and the Iberian peninsula, in three distinct evolutionarily significant units (ESUs) (Herman *et al.*, 2019). They are larger than bank voles, weighing 24-55 g and are between 9 and 13 cm in length but tend to be less aggressive (Kapusta and Sales, 2009). Predominantly solitary animals, both sexes are territorial, females particularly so whilst lactating (Loughran, 2006; Kapusta and Sales, 2009). Over winter, females exhibit avoidance behaviours and both sexes exhibit less overlap between home ranges (Agrell 1995; Agrell et al. 1996). However in spring, home ranges of males are more overlapping, and they tend to remain stationary, with regular excursions to surrounding areas (Agrell et al. 1996). Males are also highly aggressive, forming a hierarchy (Viitala, 1977; Borowski, 2003). The field vole is a polygynous species, and ovulation is induced by interaction and copulation (Breed 1967; Breed and Clarke 1970).

Comparatively little is known about communication of *M.agrestis*. Audible calls within or between sexes are important for sexual selection (Sales, Czuchnowski and Kapusta, 2007), and more time is invested into ultrasonic vocalisations than in the bank vole (Kapusta and Sales, 2009). These vocalisations are also employed by juvenile field voles, which upon separation, increase the frequency of these calls, the complexity of which increases with age (Mandelli and Sales, 2004). In terms of semiochemistry, whilst field voles have been shown to use olfactory cues to communicate with conspecifics (Stoddart, 1982), comparatively little is known regarding its composition or function.

More focus has however been on the cyclicity of field vole populations. Over a period of four years, populations decline and surge (Norrdhal 1995); the origin of these cycles is widely debated. Geographical patterns are apparent, potentially linked to the risk of predation, which varies proportionally to snow cover (Hansson and Henttonen, 1985; Hansson, 1986; Hanski, Hansson and Henttonen, 1991; Lambin, Bretagnolle and Yoccoz, 2006). The cause of population cycles appears to be incredibly complex, and associated behaviours reflect this. For example, during population increase, field voles are predominantly diurnal, however late during population decline, in July and August, field voles favour nocturnal activity (Halle and Lehmann, 1992).

Considering the complexity of field vole social structure and population variation, there is little investigation into the behaviours and communication mechanisms of this species. After investigation into the closely related *Cricetid* rodent, the bank vole (*M.glareolus*), a single sexually dimorphic,

seasonally-expressed protein was discovered in male bank vole urine that is likely to have a semiochemical function (Loxley *et al.*, 2017). The simplicity of this profile, at least at first appearance, is in stark contrast to the urinary proteome of mice and rats, which exhibit high levels of heterogeneity. The level of information required for effective olfactory communication is possibly reflective of the social structure; densely populated mice and rats rely on extensive olfactory cues to detect individual identity and kinship and prevent inbreeding. However, bank voles are predominantly solitary animals. The field vole is another rodent within the *cricetinae* family that is predominantly solitary, but with a complex cyclical population. A proteomics-based approach similar to that taken for the bank vole was taken to investigate the protein component of field vole urine and scent marks.

4.2 Aims & Objectives

The protein content of house mouse scent secretions and their functionality is well-established, with complex protein signals thought to be reflective of their dense social structure. The urinary protein output of the bank vole is both seasonal and sexually dimorphic, with mature males expressing large quantities of a single protein species during the breeding season, in contrast to female protein output and male ouput outside of the breeding season. This simple protein expression pattern of these solitary animals appears to support the link between signal complexity and social structure, however investigation of the protein output of a greater number of species is required to draw further conclusions. The field vole, *Microtus agrestis*, is another predominantly solitary rodent that exhibits complex population cycles and as a result, interrogation of the protein component of scent secretions could elucidate further protein-mediated scent signalling. The aims of this chapter therefore are as follows:

a. To assess the overall complexity of the urinary proteome of the field vole, and to identify and characterise major proteins.
b. To isolate and sequence proteins whose abundance, expression pattern or homology suggests a putative role in chemosignalling.
c. To investigate changes in presence and abundance of proteins based on sex.
d. To characterise the protein content in field vole scent marks.

4.3 Methods

4.3.1 Housing conditions

As described in section 2.3.1.

4.3.2 Urine collection

As described in section 2.3.2.

4.3.3 Scent mark collection

As described in section 2.3.3.

4.3.4 Protein assay

As described in section 2.4.1.

4.3.5 Creatinine assay

As described in section 2.4.2.

4.3.6 Polyacrylamide gel electrophoresis

As described in section 2.4.3.

4.3.7 In-gel digestion

As described in section 2.5.1

4.3.8 In-solution digestion

As described in section 2.5.2.

4.3.9 Electrospray-mass spectrometry of intact proteins

As described in section 2.6.1.

4.3.10 Tandem mass spectrometry

As described in section 2.6.3.

4.3.11 Database searching

As described in section 2.7.1.

4.3.12 De novo sequencing

As described in section 2.7.2.

4.3.13 Label-free quantification

As described in section 2.7.3.

4.3.14 Anion Exchange Chromatography

As described in section 2.9, with the following amendments.

The column (RESOURCEQ™ 1 mL, GE Healthcare) was equilibrated in 10 mM HEPES, pH 8.0 over 10 column volumes. Protein (500 µg) was loaded and eluted over a gradient of 0 – 300 mM NaCl in 10 mM HEPES pH 8.0 for 20 column volumes at a flow rate of 1 mL/min. The column was washed between runs with a 10 column volume flush of 10 mM HEPES, 1 M NaCl pH 8.0.

4.3.15 Desalting columns

As described in section 2.10.3.

4.3.16 BLAST searching

As described in section 2.11.1.

4.3.17 Multiple Sequence Alignment

As described in section 2.11.2.

4.3.18 Structural Homology Modelling

Structural models of field vole glareosin were produced using the Protein Homology/analogy Recognition Engine V 2.0 (Phyre^2) server (http://www.sbg.bio.ic.ac.uk/phyre2/html/) and modelled to the highest scoring homologous protein structure in the PDB database, selected by searching the primary sequence using BLAST against all known *Rodentia* protein structures.

4.3.19 Statistical Analysis

As described in section 2.11.5.

4.4 Results

4.4.1 Initial Assessment

Protein concentration was assessed using the Bradford assay for 95 male and 62 female urinary samples from a total of 34 mature male and 23 mature female animals kept in light cycle conditions simulating the breeding season (Figure 4.1). Creatinine levels of breeding season samples did not differ significantly between males and females ($\chi^2(1)=2.60$, $p=0.11$) (Figure 4.1A), and consequently protein output was assessed relative to creatinine output to correct for urine dilution. Protein output corrected for creatinine was higher in males than females ($\chi^2(1)=6.40$, $p=0.01$), where females had a lower protein (mg)/creatinine (mg) ratio by a value of 1.2 ± 0.5 µg/mL (mean ± SE) (Figure 4.1C). Uncorrected protein was also lower in females ($\chi^2(1)=12.96$, $p<0.01$) by 403.2 ± 113.5 µg/mL (mean ± SE)) (Figure 4.1B).

Protein content was also assessed by SDS-PAGE (Figure 4.1D) and native PAGE (Figure 4.1E). Sample nomenclature as given in Figure 4.1 for SDS-PAGE analysis correlates with nomenclature of intact protein analysis and supplementary data. Samples were corrected for urine dilution by loading 0.3 µg creatinine. Two predominant protein bands were observed in male samples at approximately 16 and 17 kDa, resolving similarly to glareosin, the abundant urinary lipocalin in the bank vole (16930 Da) (Loxley *et al*., 2017). Fainter bands were also observed at approximately 21 and 22 kDa, the latter of which was also commonly seen in females.

Urinary proteins of 13 male and 5 female samples were analysed by ESI-MS. Male samples were diluted in 95% HPLC-grade-H_2O, 5% MeCN, 0.1% formic acid and 1 pmol protein (assuming a predominant protein Mr of 16 kDa based on SDS-PAGE) was analysed. Female samples were desalted using Zeba™ Spin Desalting Columns (Thermo Scientific) before dilution in the same buffer and 10 pmol injected into the instrument. Acquired spectra were deconvoluted over a large mass range, to identify approximate masses of main protein species. Once an approximate mass had been established, deconvolution was applied over a small mass range with a smaller mass error applied (Figure 4.2; Figure 4.3).

Four predominant masses were recurring in males; 17138, 17168, 17236 and 17252 Da (spectrum from pooled male samples, Figure 4.4). Other masses recurring include 16893 Da in addition to a number of masses approximating 21 kDa at low levels. In contrast, no dominant peaks at approximately 17 kDa were observed in female profiles, however, the cluster of masses at 21 kDa dominated the spectrum, in addition to a singular lower mass of 16648 Da (spectrum from pooled female samples, Figure 4.4). In female urine samples, masses observed included 21440 ± 1 and 21831 ± 1 Da, which have a mass difference of 307 Da that could possibly arise from a modification

of N-glycolyl neuraminic acid, a sialic acid commonly occurring at the end of carbohydrate chains of glycoproteins. Two peaks were also observed at 21483 ± 1 and 21831 ± 1 Da, which correspond to mass increments of 42 Da relative to 21440 ± 1 and 21831 ± 1 Da, respectively. Acetylation of a protein or peptide results in an increase in molecular weight by 42 Da, so could be a suitable explanation for this mass discrepancy. Acetylation is a common modification of glycan chains, and combined with evidence of a mass increment corresponding to a sialic acid, it is suggested that the cluster of masses of approximately 21 kDa could be a result of polymorphism at the modification level.

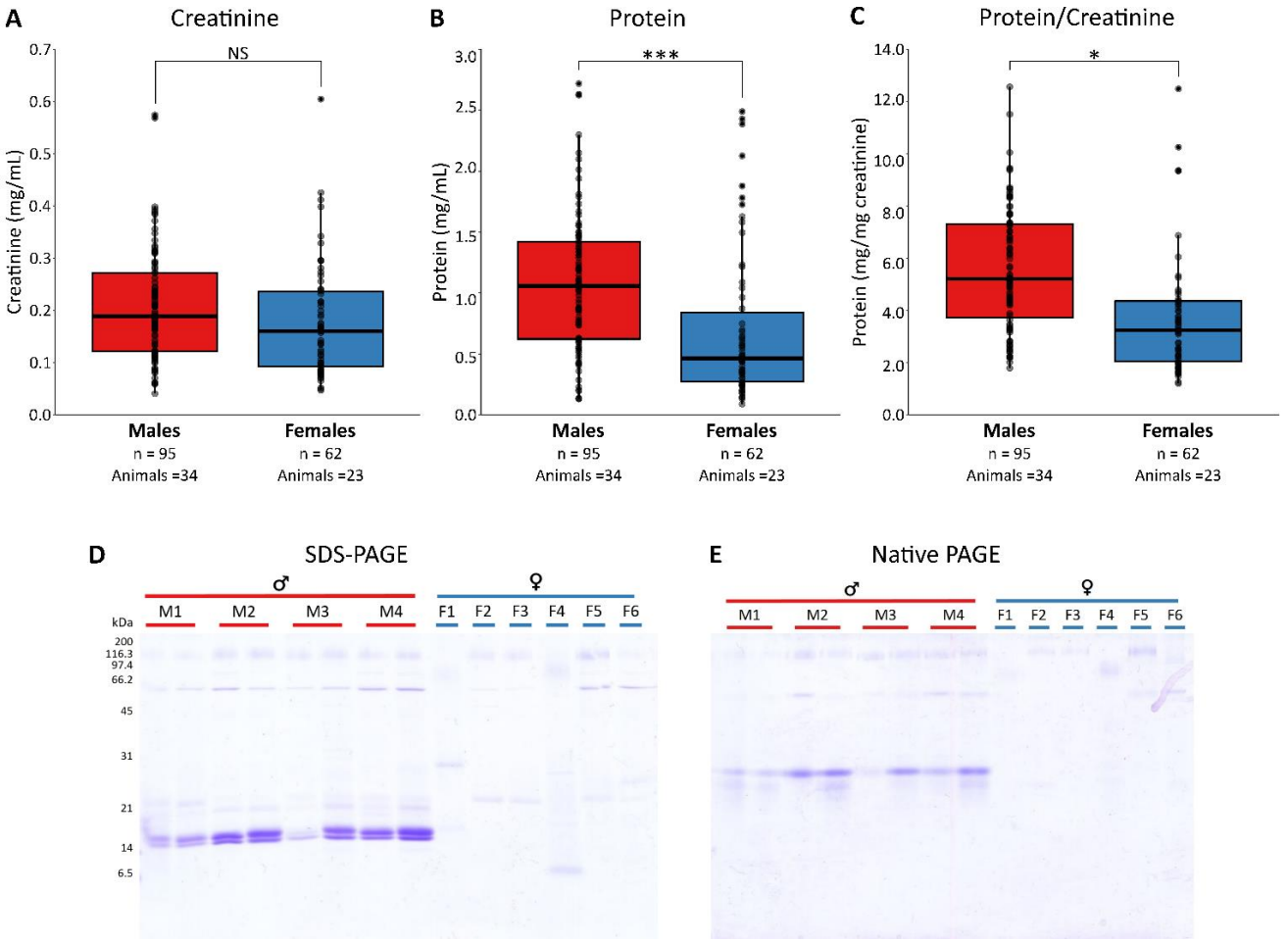


Figure 4.1 | Analysis of the protein output of field vole urine samples.
The overall protein output of male (n = 95) and female (n = 62) urine samples from 34 male and 23 female field voles was assessed. No statistically difference was observed in urine dilution (A), measured by creatinine output. Protein output (B) was significantly higher in male samples (*** = $p < 0.01$) and protein corrected for urine dilution (C) was also significantly increased in males (* = $p < 0.05$). Samples were corrected for urine dilution by loading 0.3 µg creatinine before SDS-PAGE (D) and native PAGE (E) analysis. Samples are labelled according to individual animal from which the urine samples were obtained. Multiple gel lanes from the same individual are indicative of different urine samples taken at different times.

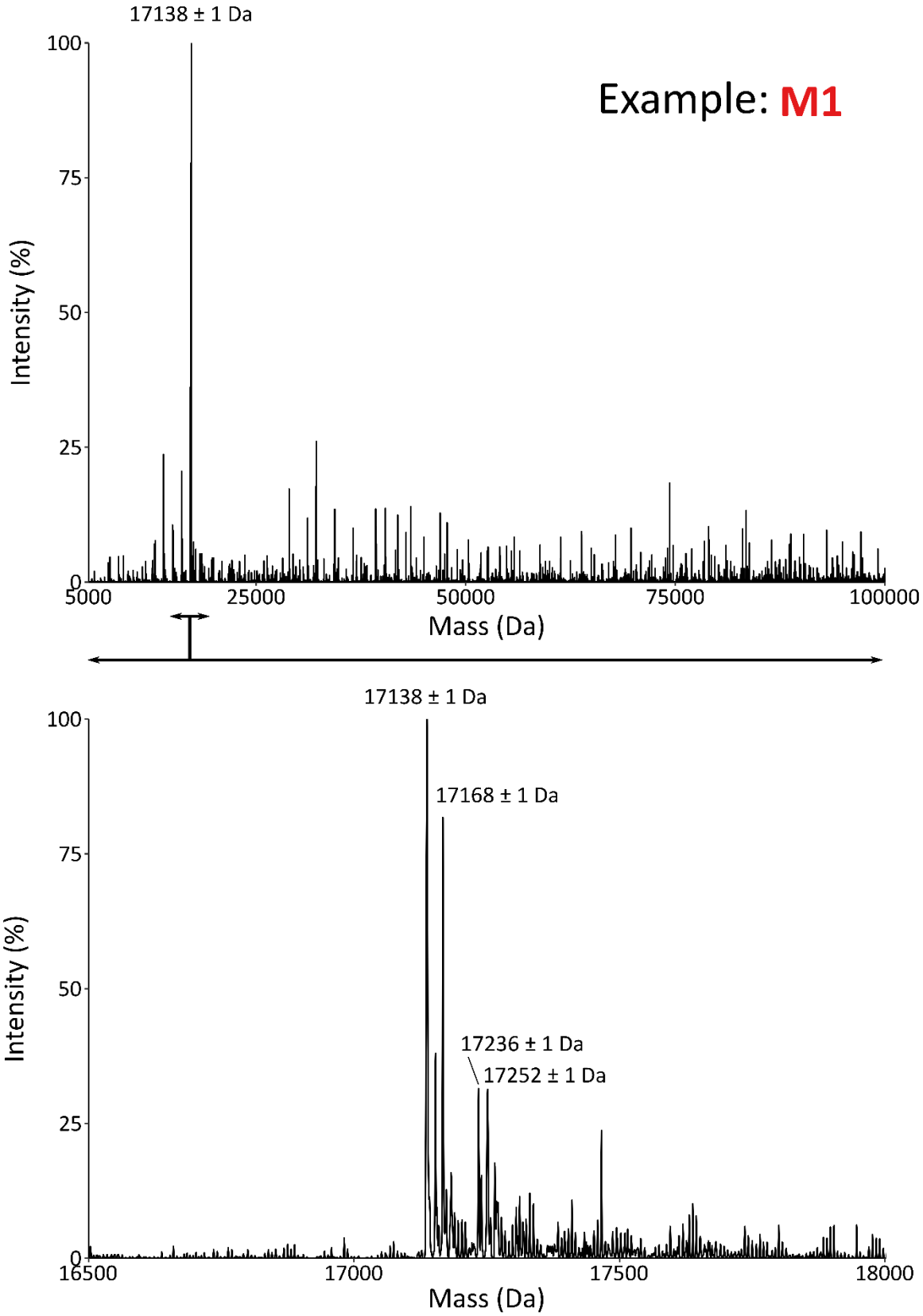


Figure 4.2 | ***Intact mass analysis of male field vole urinary proteins.***
Male field vole urine samples (n = 10) (1 pmol) were analysed by ESI-MS. Spectra were deconvoluted within a larger mass range (above) to determine an approximate mass of the predominant proteins. Once the approximate mass had been established, spectra were deconvoluted over a smaller mass range (5000 – 30000 Da). Example of a single sample (M1; lane 1 PAGE analysis).

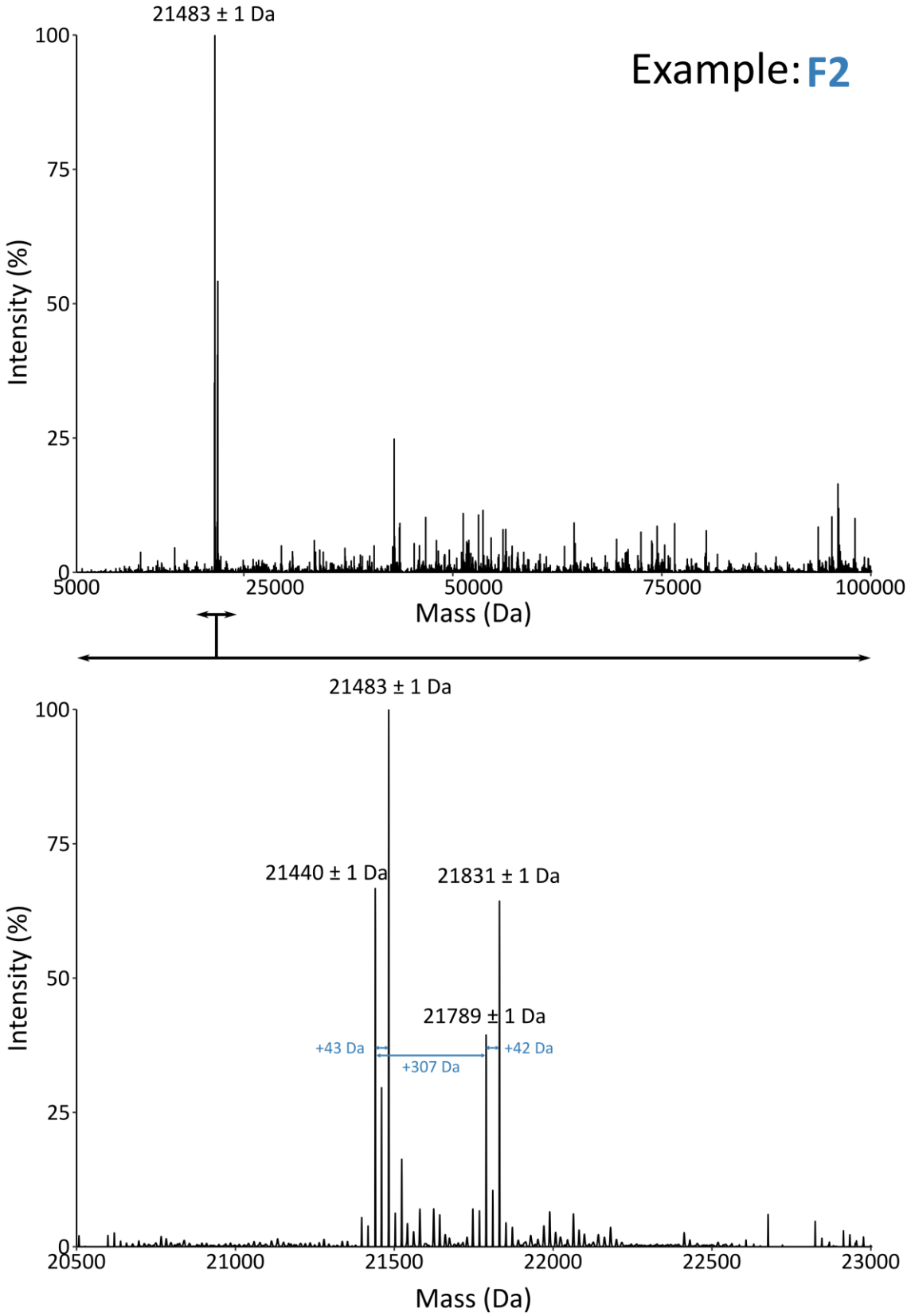


Figure 4.3 | *Intact mass analysis of female field vole urinary proteins*
Female field vole urine samples (n = 6) (10 pmol) were analysed by ESI-MS. Spectra were deconvoluted within a larger mass range (above) to predict an approximate mass of the predominant proteins. Once the approximate mass had been established, spectra were deconvoluted over a smaller mass range (5000 – 30000 Da). Example of a single sample (F2; PAGE analysis).

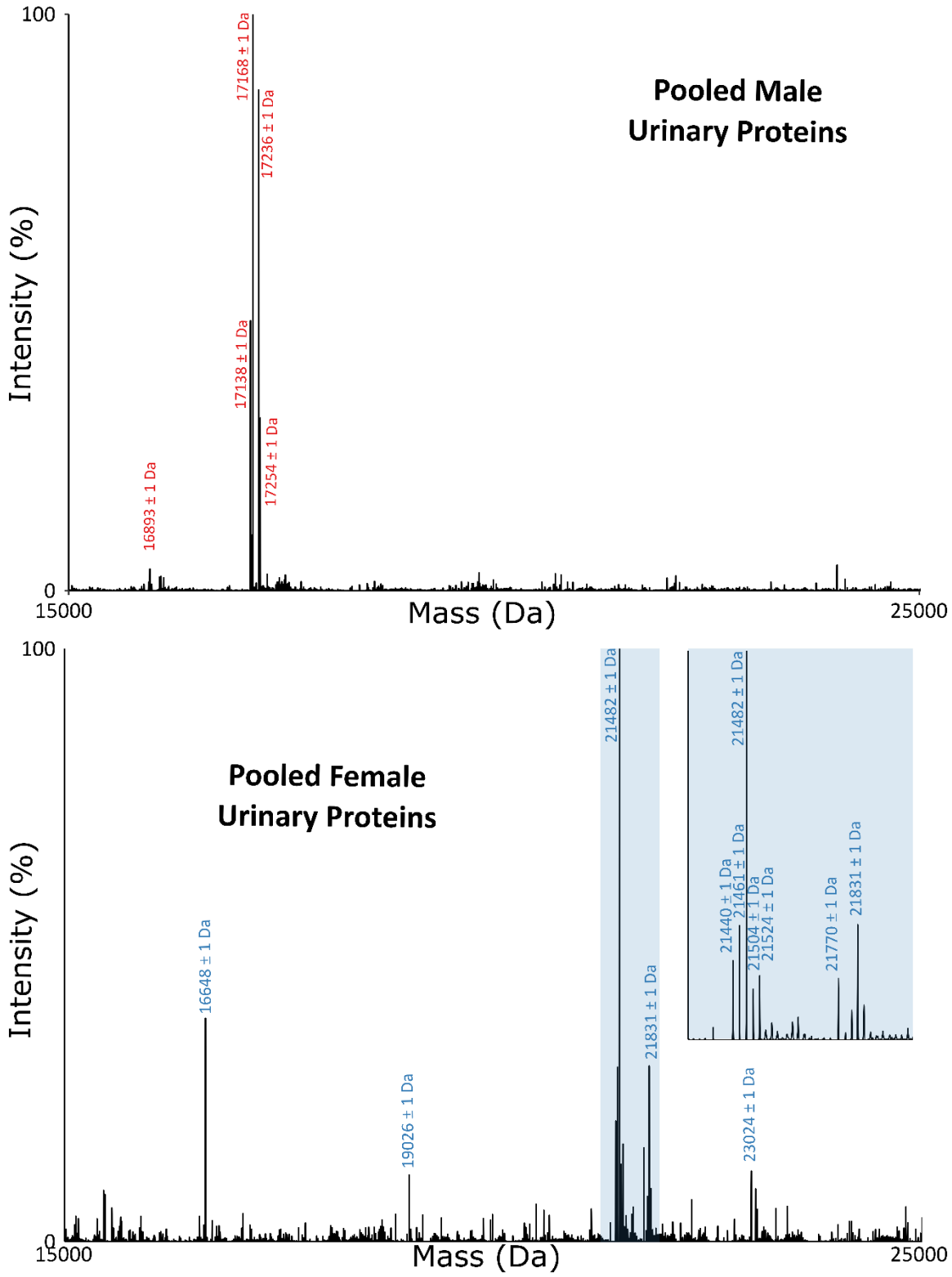


Figure 4.4 | Intact mass profiling of pooled male and female field vole urine samples.
Intact proteins in pooled urine from 13 male field voles (1 pmol), top, and 5 female field voles (10 pmol, desalted), bottom, were analysed by ESI-MS. Spectra were deconvoluted within a larger mass range (above) to predict an approximate mass of the predominant proteins. Once the approximate mass had been established, spectra were deconvoluted over a smaller mass range (5000 – 30000 Da).

4.4.2 Identification of major male field vole urinary proteins

LC-MS/MS analysis of proteins digested in-solution from pooled male urine provided strong evidence of peptides homologous to both glareosin and odorant binding proteins (OBPs) from the bank vole and an aphrodisin-like protein in the Chinese hamster, *Cricetulus griseus* (UniProt accession A0A061HV58) (Figure 4.5). They were among the highest-scoring proteins identified, which included albumin, uromodulin and serotransferrin from numerous species in the UniProt database. Also identified were major urinary protein-like sequences from the golden hamster, *Mesocricetus auratus* (Figure 4.6). Inspection of peptide LC peak areas suggested that a protein homologous to bank vole glareosin was the highest abundant identified protein and manual inspection of the highest abundant unidentified spectra sequenced *de novo* by PEAKS also revealed many OBP-like sequences. It was therefore likely that the abundant masses observed from analysis of intact proteins in pooled male urine of approximately 17 kDa were most closely related to bank vole glareosin and OBPs.

To gain further sequence information, an alignment of the bank vole glareosin and OBP sequences was used as a scaffold on which to produce an estimate of a field vole glareosin consensus sequence. Peptide matches and *de novo* tags indicated in the PEAKS^{TM} peptide map in Figure 4.5 were manually inspected and good quality sequences from high intensity signals were selected, and in combination with high-scoring unidentified spectra sequenced *de novo*, a consensus protein sequence was built from overlapping peptides (Figure 4.7A). The sequence aligned well with bank vole glareosin and OBPs, and contained both the conserved lipocalin motif [GXW] in addition to four cysteine residues that align perfectly with those of the other sequences and consequently suggest the presence of 2 disulfide bonds, as indicated (Figure 4.7). Data were re-searched against the sequence with strong peptide evidence and 100% sequence coverage (Figure 4.8).

Peptides from male field vole urine mapped by cross-species matching to known lipocalin sequences

Glareosin (*M.glareolus*)

-10lgP = 158.65

Sequence coverage = 68%

OBP3 (*M.glareolus*)

-10lgP = 184.74

Sequence coverage = 22%

Identified peptide *De novo* tag

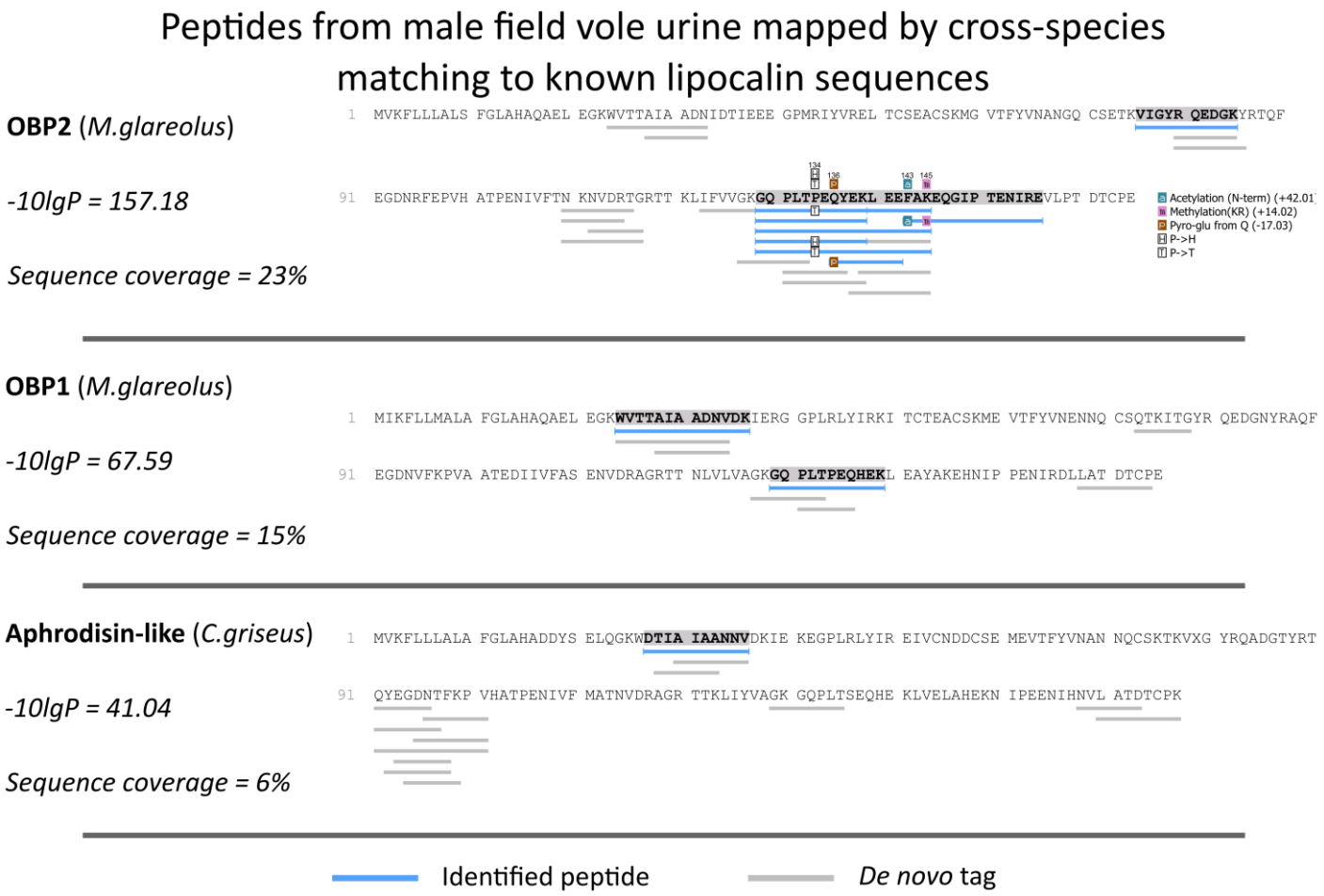


Figure 4.5 | Peptide coverage of glareosin and the odorant binding proteins (OBPs) when searching peptides generated from proteolytic digestion of proteins in pooled male field vole urine.

Urine from male field voles was pooled and subject to digestion with four different proteolytic enzymes. Resulting peptides were analysed by LC-MS/MS using an Orbitrap mass spectrometer (Thermo Q Exactive™). Data were inspected in PEAKS™ (Bioinformatics Solutions Inc.) and searched against a database comprising all UniProt sequences within *Rodentia* in addition to the previously established bank vole glareosin protein. OBP-like sequences, including bank vole glareosin, were confidently identified.

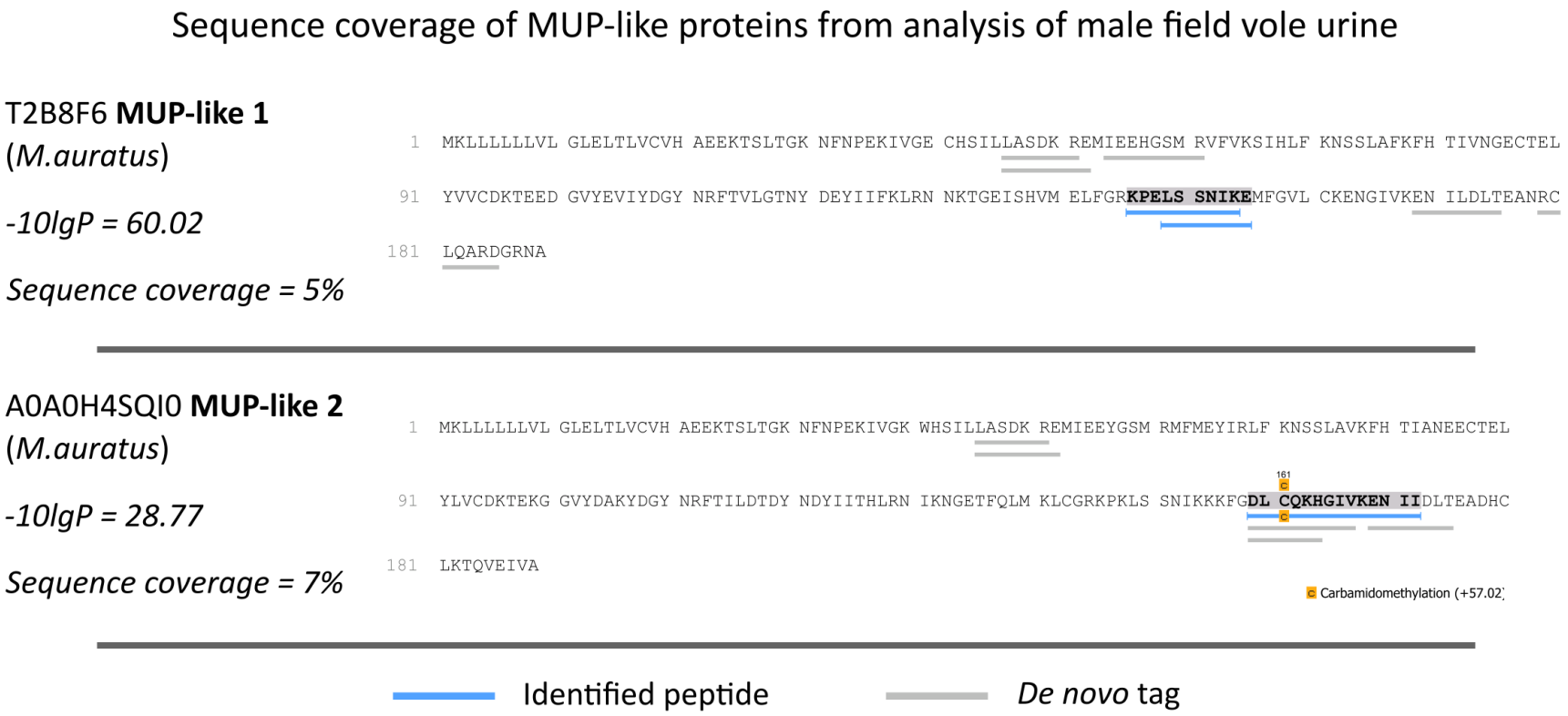


Figure 4.6 | Peptide coverage of major urinary protein-like sequences from the golden hamster when searching peptides generated from proteolytic digestion of proteins in pooled male field vole urine.
Urine from male field voles was pooled and subject to digestion with four different proteolytic enzymes. Resulting peptides were analysed by LC-MS/MS using an orbitrap mass spectrometer (Thermo Q Exactive™). Data were inspected in PEAKS™ (Bioinformatics Solutions Inc.) and searched against a database comprising all UniProt sequences within *Rodentia* in addition to the previously established bank vole glareosin protein. MUP-like sequences from the golden hamster, *Mesocricetus auratus*, were confidently identified.

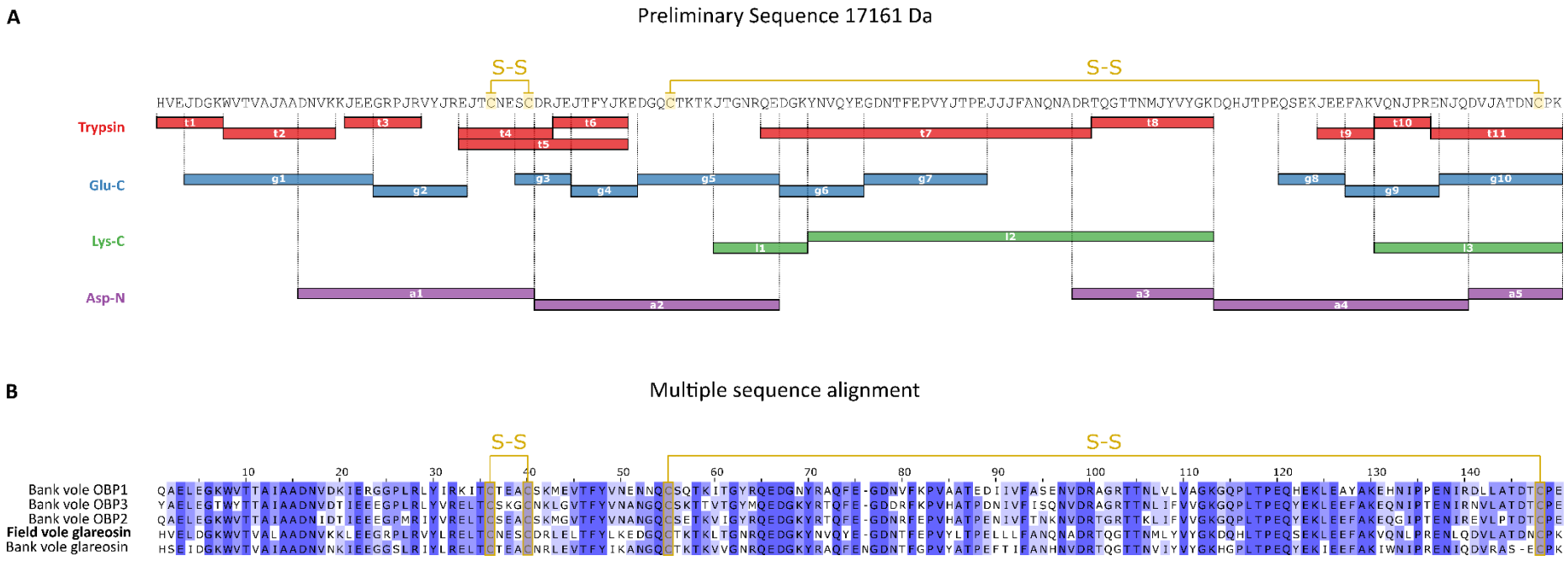


Figure 4.7 | Preliminary field vole glareosin sequence built from overlapping peptides generated in PEAKS™.
Overlapping peptides, generated from the LC-MS/MS analysis of multiple proteolytic digests of pooled male field vole urine, were used to construct an initial version of field vole glareosin (A). The letter 'J' is used to indicate the position of either leucine or isoleucine. Peptides from proteolytic cleavage using trypsin (t, red), glu-C (g, blue), lys-C (l, green) and asp-N (a, purple) were used to assemble the protein sequence. Cysteine residues are highlighted in yellow, and labelled as disulfide bond pairs in line with homologous pairs from bank vole glareosin and other OBPs. The sequence was then aligned with bank vole glareosin and OBPs using Clustal Omega (Sievers *et al.*, 2011) and visualised in Jalview (Waterhouse *et al.*, 2009) (B).

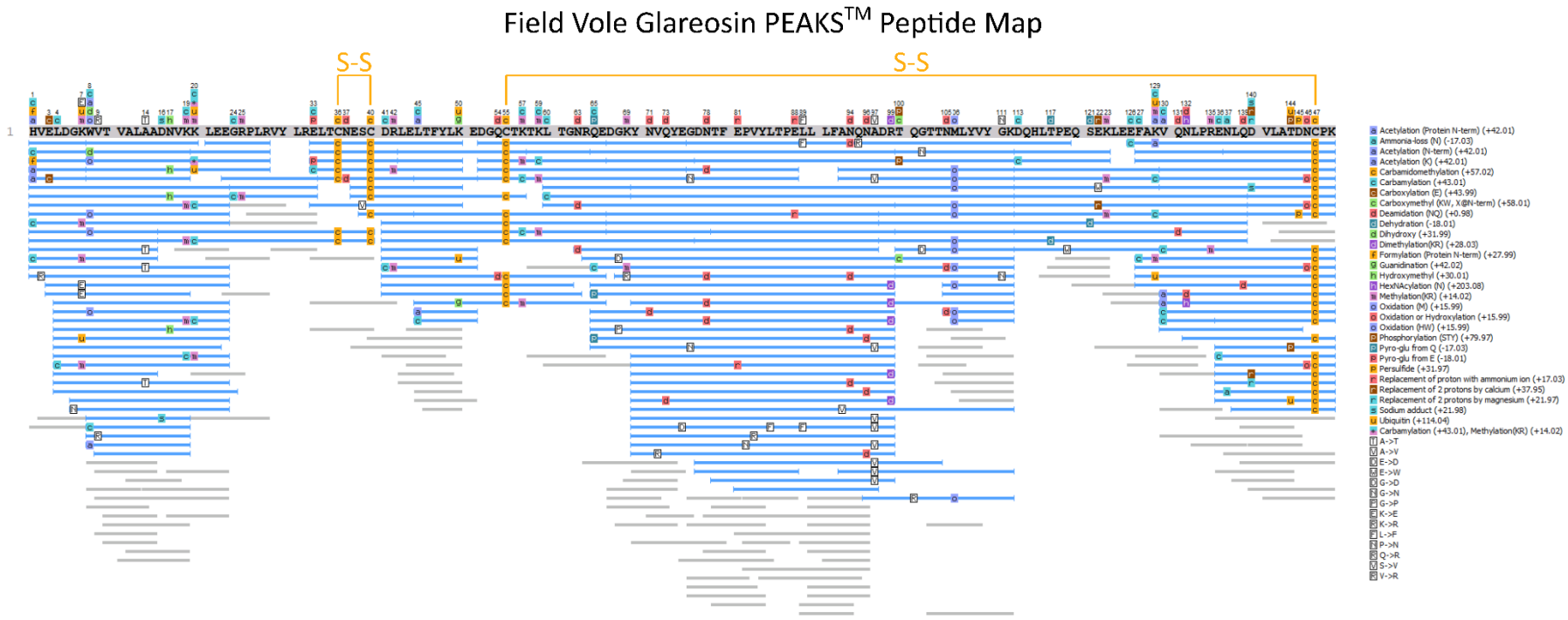


Figure 4.8 | *Peptide coverage of initial field vole glareosin sequence.*
Peptides generated from four different proteolytic digestions of urinary proteins from pooled male urine were analysed by LC-MS/MS and data were examined in PEAKS™ by searching against the novel 'scaffold' protein sequence.

The ‘scaffold’ sequence generated has a calculated average mass of 17161 Da. Allowing for a 4 Da loss from the formation of two disulfide bonds, this would have an observed mass of 17157 ± 1 Da, which did not correspond to any of the masses observed from analysis of intact urinary proteins. The five predominant masses observed from intact protein analysis were 16983, 17138, 17168, 17236 and 17252 ± 1 Da. Of the mass differences (155, 30, 68 and 16 Da), only the 16 Da is one which corresponds to a commonly seen modification of oxidation. To explore the possibility that this heterogeneity is a result of non-synonymous point mutations, closer inspection of the PEAKS™ results sequence coverage map from a SPIDER search reveals a number of peptides with suggested mutations (Figure 4.8).

Mutations suggested by the software, in addition to the *de novo* tags identified in the search against the ‘scaffold’ field vole glareosin sequence, were closely inspected and the MS2 spectra were manually assessed. A possible 15 mutations were found at 10 different sites in four peptides (Figure 4.9),allowing a possible 96 combinations and consequently 96 possible total masses. Of these, 7 combinations resulted in calculated total masses matching those observed in intact mass spectra, allowing for the probable loss of 4 Da in the intact mass spectra due to the conserved cysteine positions (the alignment of which would not have been affected by any of the mutations found) (Table 4.1).

One combination was attributable to the observed mass 17138 Da (-19 Da), which changed only one peptide from the scaffold sequence, from DQHJTPEQSEK > GQPJTTEQYEK, a loss of 18 Da. Two combinations were calculated as possible origins of the observed mass 17168 Da (+11 Da). Both contained the 18 Da loss, with the remaining +30 Da difference possibly attributable to a point mutation either in the peptide WVTVAJ[A>T]ADNVK (+ 30 Da) or in the peptide YNVQYEGDNTFEPVYJTPEJJJFANQN[A>V]DR (+28 Da), both of which are feasible given the mass error limits of the instrument used for intact mass profiling. For the observed mass at 17236 Da (+79 Da), two combinations were also considered. Mutations in the peptide YNVQYEGDNTFEPVY[J>M]TPE[J>F]JJFANQN[A>V]DR (+80 Da), or alternatively a combination of W[V>TA]TVAJAADNVK (+73 Da), YNVQYEGDNTFEPVYJTPEJJJFANQN[A>V]DR (+28 Da) and GQPJTTEQYEK (-18 Da) could cause the mass shift. For the final predominant mass peak in intact mass analysis, 17252 Da (+95 Da), this could have been caused by JTC[N>K]ESCDRJE (+14 Da) and YNVQYEGDNTFEPVY[J>M]TPE[J>F]JJFANQN[A>V]DR (+ 80 Da) in regards to the scaffold protein. Alternatively, the mass peak 17252 ± 1 Da could be the result of oxidation, as a mass increment of 16 Da, the difference between 17236 and 17252 Da, is indication of this common modification.

Whilst there is undoubtedly peptide evidence with considerable similarity to bank vole glareosin, the origin of the observed heterogeneity required further experimental assessment. To assist with identifying the cause of heterogeneity, anion exchange chromatography was used to further separate proteins prior to sequencing *de novo*.

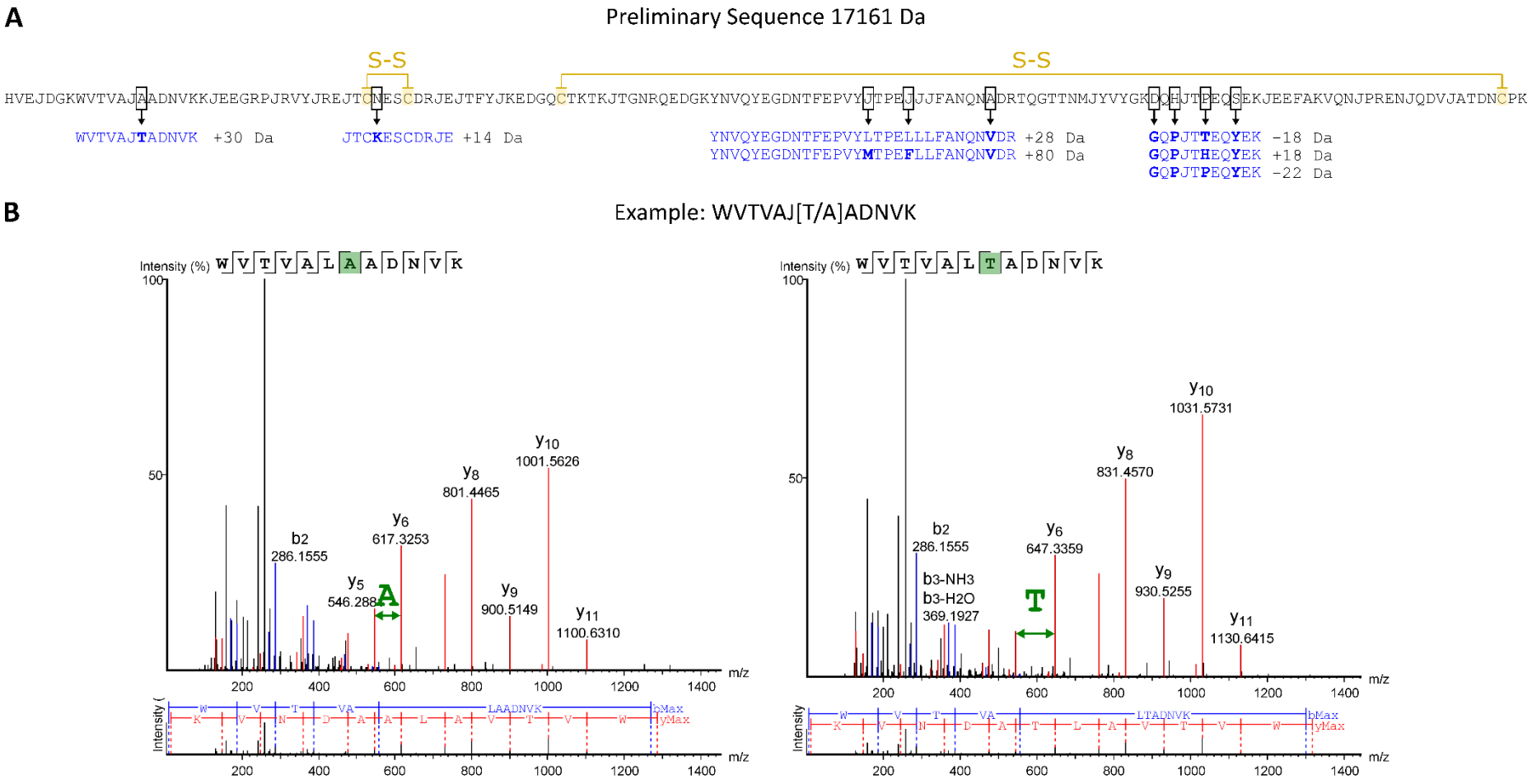


Figure 4.9 | *Sites of possible heterogeneity in field vole glareosin.*
LC-MS/MS data from multiple enzymatic digestion of pooled male urinary proteins were searched against an initial scaffold sequence for field vole glareosin. *De novo* tags and suggested mutations were investigated to identify possible heterogeneity that would explain a complex intact mass profile (A). MS2 spectra for each possible mutation was manually inspected (B).

Table 4.1 | *Possible mutation combinations of field vole glareosin.*

Possible heterogeneity in field vole glareosin was explored by examination of 15 mutation sites identified by SPIDER searched in PEAKS^{TM} or inspection of *de novo* tags. Mutation combinations (n=5) correspond to observed masses in intact mass profiles. Amino acids highlighted in red indicate putative mutation sites.

Observed Mass (±1 Da)	**Mass after reduction of disulfide bonds (±1 Da)**	**Possible Mutation Combination**				
		Mass (±1 Da)	*Peptide 1*	*Peptide 2*	*Peptide 3*	*Peptide 4*
17138	17142	17143	WVTVAJ**A**ADNVK	JTC**N**ESCDRJE	YNVQYEGDNTFEPVY**J**TPE**J**JJFANQN**A**DR	**G**Q**P**JT**T**EQ**Y**EK
17168	17172	17171	WVTVAJ**A**ADNVK	JTC**N**ESCDRJE	YNVQYEGDNTFEPVY**J**TPE**J**JJFANQN**V**DR	**G**Q**P**JT**T**EQ**Y**EK
		17173	WVTVAJ**T**ADNVK	JTC**N**ESCDRJE	YNVQYEGDNTFEPVY**J**TPE**J**JJFANQN**A**DR	**G**Q**P**JT**T**EQ**Y**EK
17236	17240	17241	WVTVAJ**A**ADNVK	JTC**N**ESCDRJE	YNVQYEGDNTFEPVY**M**TPE**F**JJFANQN**V**DR	**D**Q**H**JT**P**EQ**S**EK
17252	17256	17255	WVTVAJ**A**ADNVK	JTC**K**ESCDRJE	YNVQYEGDNTFEPVY**M**TPE**F**JJFANQN**V**DR	**D**Q**H**JT**P**EQ**S**EK
16893	16897	N/A	No combinations below 17 kDa			

4.4.3 Primary sequence-level heterogeneity of field vole glareosin

Pooled urine from male field voles was desalted on Zeba™ desalting columns (Thermo) and protein (500 µg) was loaded onto a 1 mL ResourceQ™ anion exchange column attached within an ÄKTA Purifier off-line chromatography system. Proteins were eluted off the column using a gradient of 0 – 300 mM NaCl in 10 mM HEPES pH 8.0 over 20 minutes and monitored with UV absorbance at 280 nm. Fractions were collected manually and the protein content was assessed by Bradford assay and SDS-PAGE (10 µL) (Figure 4.10). Fraction 9 was the predominant protein-containing fraction at almost 80 µg/mL, and giving a clear protein band at approximately 17 kDa on SDS-PAGE. However, fractions 8, 10, 11 and 12 also contained lower levels of this protein, and these protein-containing fractions were subsequently analysed by ESI-MS (500 fmol) to assess heterogeneity (Figure 4.11).

Whilst fraction 9 contained the same protein masses in an identical ratio to those in the original pooled male urine, fraction 8 appeared to be comprised predominantly of the 17168 Da peak, whereas the spectrum for fraction 10 was dominated by the mass at 17236 Da, followed by the mass of 17252 Da, possibly a result of oxidation (mass difference 16 Da). Intact mass analysis of fraction 11 unfortunately did not yield a spectrum possible to deconvolute, but fraction 12 appears to predominantly contain the proteoform of mass 16893 Da.

To investigate the origin of these masses, proteins in each fraction were captured on Strataclean resin™ (Thermo) and subject to proteolysis with enzymes of one of four different specificities; trypsin, glu-C, lys-C and asp-N to generate overlapping peptides. Peptides were analysed by LC-MS/MS and resulting data were analysed in PEAKS™. SPIDER searches were performed against a database of all *Rodentia* sequences in UniProt, in addition to the scaffold field vole glareosin. Searches were also made against a database constructed of all peptide variations as displayed in Figure 4.9, including amino acids either side of each peptide to allow for different cleavage specificities, to assess the prevalence of each peptide variant in each fraction as a possible explanation for the intact mass heterogeneity.

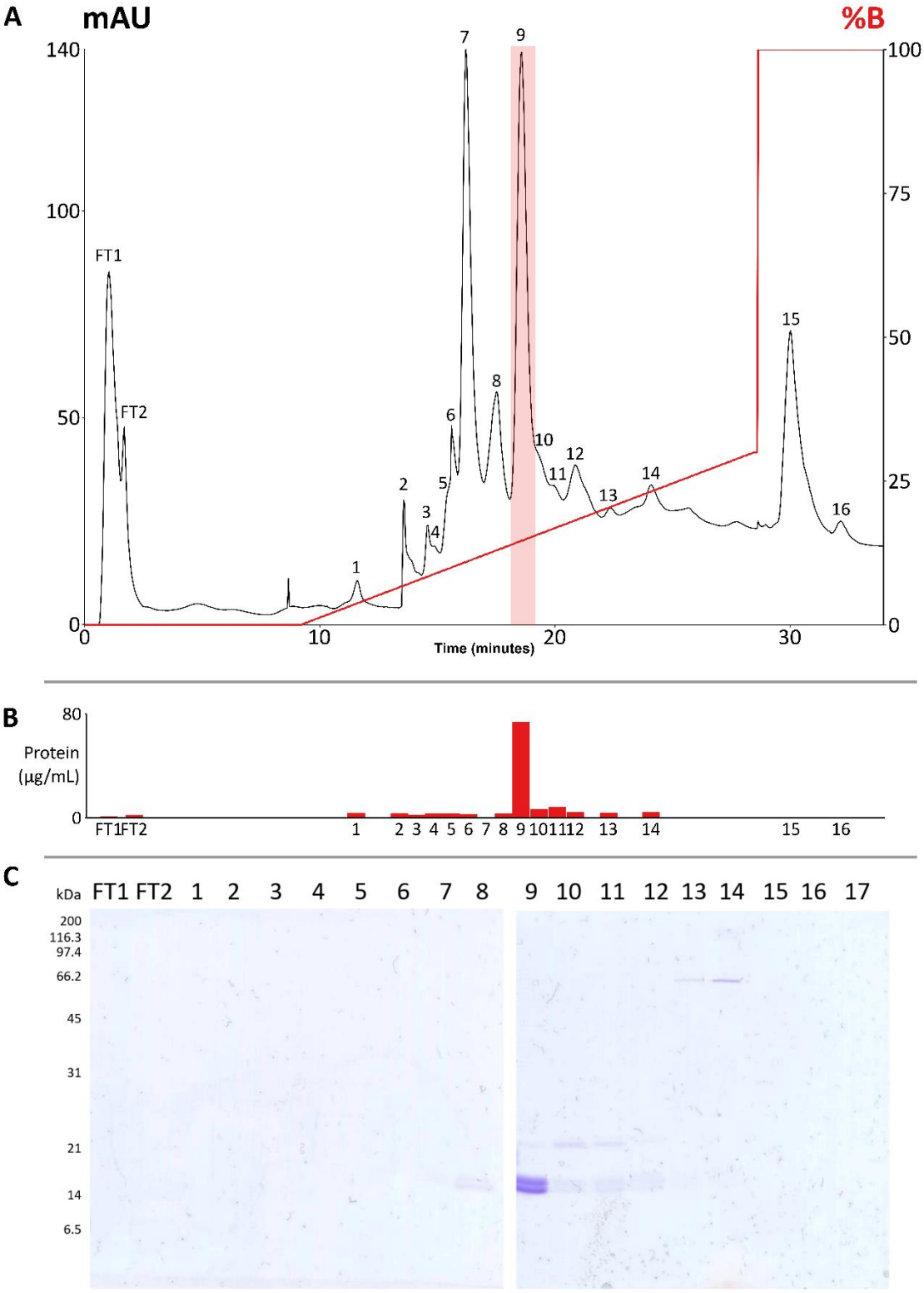
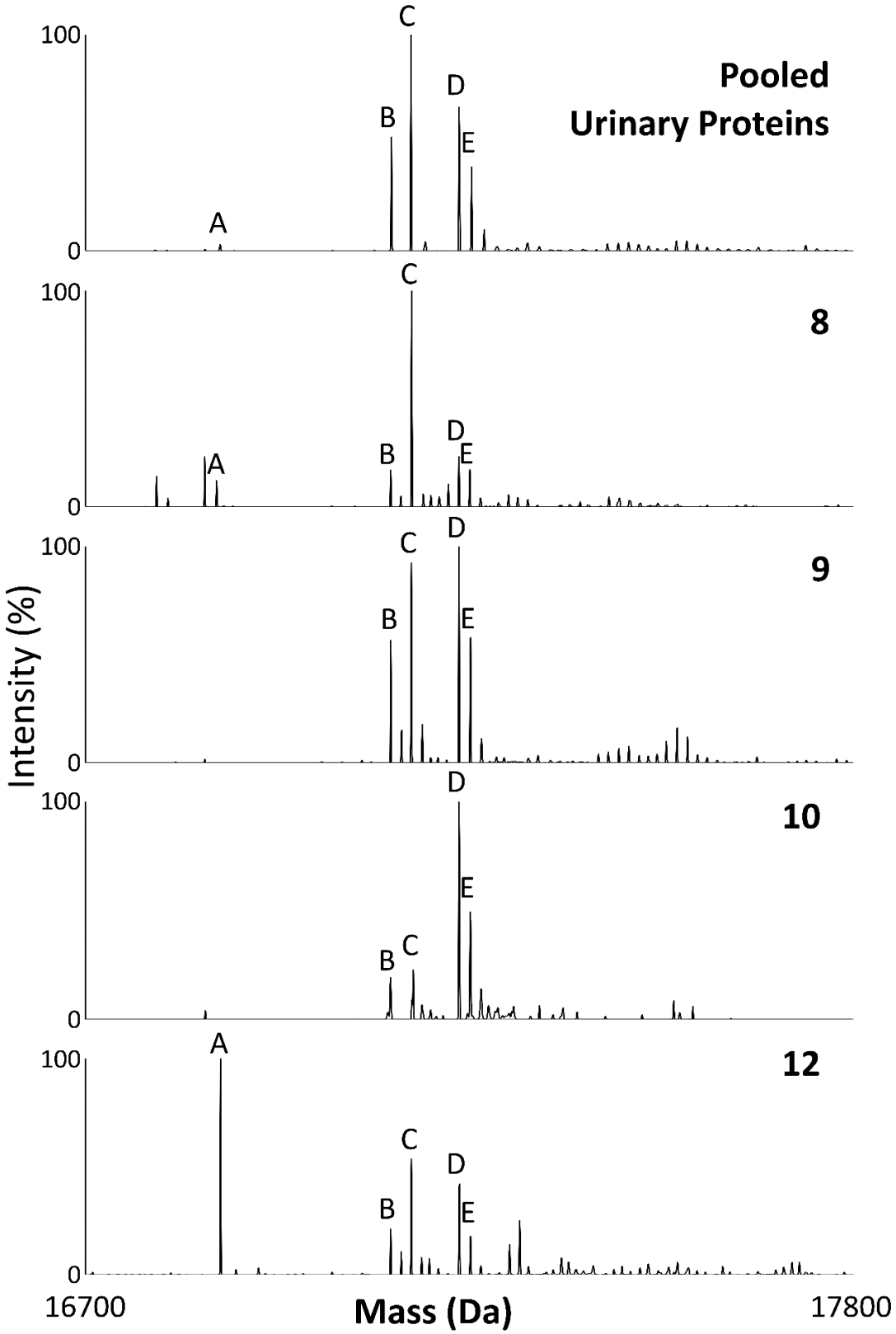


Figure 4.10 | Separation of proteins from pooled male field vole urine by anion exchange chromatography.

Proteins (500 µg) from pooled urine from sexually mature male field voles were diluted in 10 mM HEPES, pH 8.0, and loaded onto a ResourceQ™ chromatography and eluted using a gradient of 0.0 – 0.3 M NaCl. Fractions were collected manually (FT = flowthrough) by monitoring absorbance at 280 nm (A). Protein concentration was ascertained and 10 µL each fraction analysed by SDS-PAGE.



A 16893 ± 1 Da **B** 17138 ± 1 Da **C** 17168 ± 1 Da **D** 17236 ± 1 Da **E** 17252 ± 1 Da

Figure 4.11 | *Intact mass analysis of anion exchange separated proteins from pooled male field vole urine.*

Field vole urinary proteins separated by anion exchange chromatography were analysed by ESI-MS (500 fmol, based on approximate mass of 17 kDa).

LC-MS/MS analysis of peptides generated from enzymes with multiple protease specificity did not yield enough coverage to fully confirm the presence of the mutation combinations listed above in Table 4.1. This is likely due to the low concentration in the fractions of higher purity, 8 and 10, which were most important in this approach. Protein separation by anion exchange chromatography was not further pursued, in the interest of time and sample preservation, as an alternative approach was taken to analyse individual male samples with high abundance of one particular mass.

The elution of these proteins in anion exchange chromatography does however provide us with evidence that these masses are extremely similar in pI, eluting across one major peak only. If these masses were from considerably divergent sequences despite maintaining a similar mass, their elution profile would likely be further separated. The close elution of all five major masses supports the theory that they are highly similar proteins that differ in ways that do not dramatically affect their pI, such as small post-translational modifications or point mutations.

An additional resource is the order in which proteins were eluted. The proteins eluting earlier would be less negatively charged at pH 8.0 (17168 Da), and the proteins eluting later would be more negatively charged (17236 and 17252, and later 16893 Da).

An alternative approach was taken to confirm the presence of hypothesised point mutations. Whilst many protein profiles from intact mass analysis of individual male urine samples (see supplementary material) showed comparable ratios to pooled male urine (Figure 4.4), male urine samples from two individual were dominated by the mass peak 17168 Da (sample IDs 50693 & 50695) (Figure 4.12c&d). In another sample, the peak 17138 Da was highest in abundance (Male 1, lane 1 of Figure 4.1) (Figure 4.12a), and another had the highest relative abundance of 17136 Da compared to other samples (sample ID 50692) (Figure 4.12b). Proteins in each sample were digested with trypsin, lys-C, glu-C or asp-N to generate overlapping peptides to be analysed by LC-MS/MS. Data were analysed in PEAKS^{TM} by searching against a database comprising the individual mutated peptides specified in Table 4.1, flanked by additional amino acids from the scaffold protein to assist with identification in peptides from multiple proteases.

The two male samples dominated by the mass of 17168 Da provided strongest evidence for the peptides WVTVAJAADNVK, EJTCNESCDRJE, YNVQYEGDNTFEPVYJTPEJJJFANQNVDR and GQPJTTEQYEK, consistent with a protein of total mass 17171 Da and therefore a likely candidate for the peak mass 17168 Da, allowing for disulfide bond formation and instrument error tolerance. Two male samples with more complex protein profiles provided evidence of the above peptides, in addition to the peptides EJTCKESCDRJE, YNVQYEGDNTFEPVYJTPEJJJFANQNADR, YNVQYEGDNTFEPVYMTPEFJJFANQNVDR and DQHPJTPEQYEK. The peptide data support the

proposed mutations in Table 4.1, in which the mass difference between the peaks of 17138 and 17168 ± 1 Da is due to the mutation YNVQYEGDNTFEPVYJTPEJJJFANQN[A>V]DR, causing a mass increase of 28 Da. Evidence of the alternative mutation proposed to account for this mass, WVTVAJ[A>T]ADNVK was only seen at small levels in samples 50689 and 50692, but in neither of the samples that were dominated solely by the peak 17168 ± 1 Da. When searched against the data from fraction 8 (enriched in mass 17168 Da), this lack of evidence was consistent, with good sequence coverage ofthe peptide WVTVALAADNVK. However the 91A>V mutation could not be confirmed using the fractionated protein due to a lack of sequence coverage in the central region of the protein.

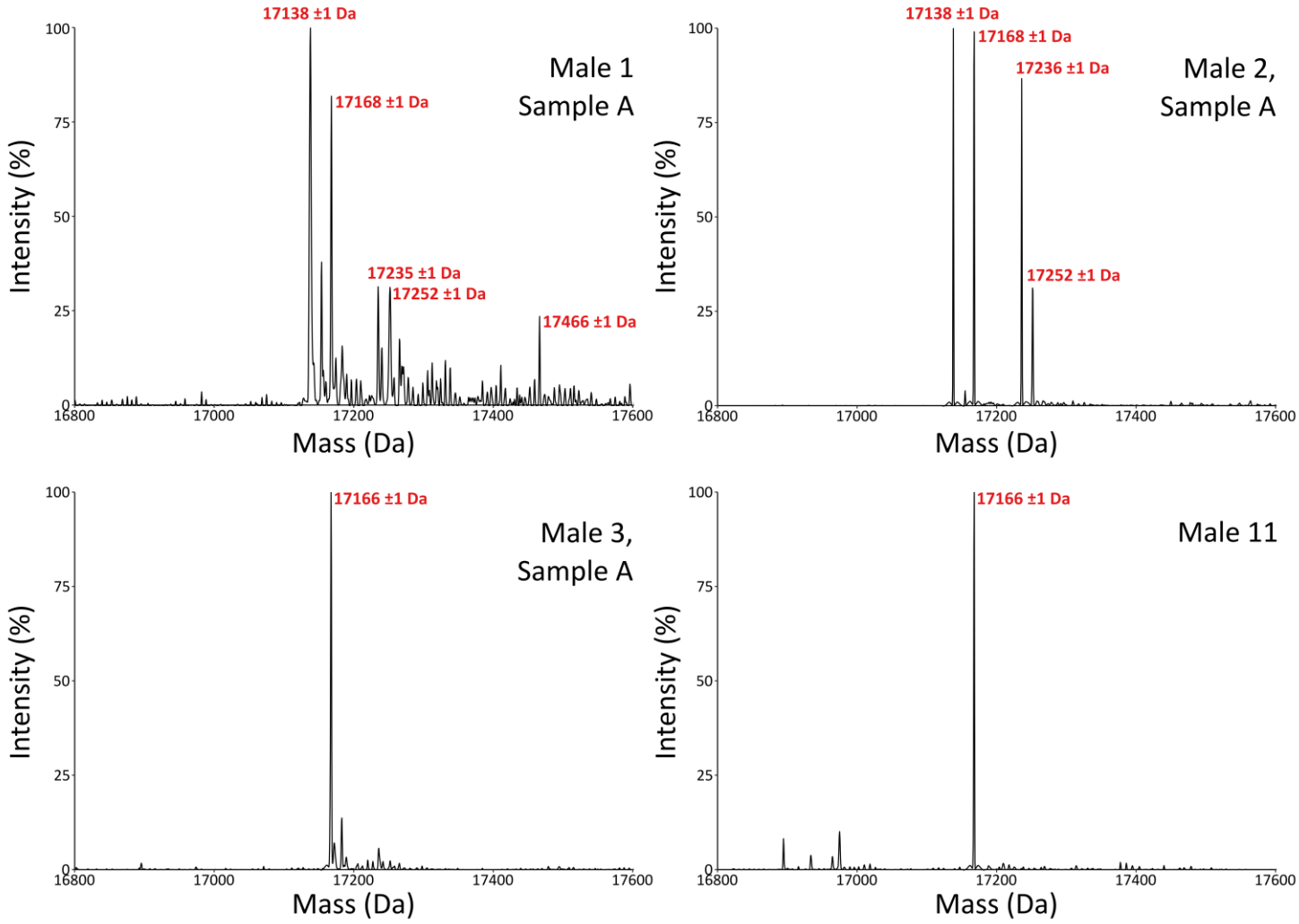


Figure 4.12 | Protein profiles of male field vole urine samples used to confirm presence of point mutations.
Four male field vole urine samples were selected for LC-MS/MS analysis after digestion with multiple proteases based on their protein profiles generated from intact mass analysis.

The four novel sequences described in Table 4.1, selecting the mutation determined above for the mass 17168 Da (A91>V), were incorporated into a database of all *Rodentia* sequences in UniProt, including bank vole glareosin. For each individual sample, the peptides from digests of different specificity were collectively searched and matches to each novel protein sequence, were inspected. For both 50693 and 50695, the two profiles dominated by the peak for 17168 ± 1 Da, only protein sequences of overall mass 17171 and 17255 Da (hypothetically corresponding to intact mass peaks of 17168 and 17252 ±1 Da, respectively) were identified (Figure 4.13). When comparing sequence coverage for these two protein sequences, sequence coverage of the YNVQYEGDNTFEPVYJTPEJJJFANQN**V**DR peptide (from sequence corresponding to 17168 Da peak) was much stronger in comparison to the YNVQYEGDNTFEPVYJTPEJJJFANQN**A**DR peptide (from sequence corresponding to 17138 Da peak), which was not identified, and the YNVQYEGDNTFEPVY**M**TPE**F**JJFANQN**V**DR peptide (from sequence corresponding to 17236 or 17252 Da peaks)(Figure 4.13, sample 50693; supplementary material, sample 50695).

In samples50689 and 50692, the proteoform with the highest score in both cases had a total mass of 17143 Da, corresponding to the intact mass peak of 17138 ± 1 Da. This is consistent with the protein profiles. The other two proteoforms identified, with masses 17171 and 17255 Da had comparable scores, but together demonstrated evidence for the peptides YNVQYEGDNTFEPVYJTPEJJJFANQN**V**DR, YNVQYEGDNTFEPVY**M**TPE**F**JJFANQN**V**DR and EJTC**K**ESCDRJE (Figure 4.14). Unfortunately, the proposed sequence for the intact mass peak of 17236 Da (total mass 17240 Da) did not have any unique peptides with which to identify it; the peptide EJTC**N**ESCDRJE is shared with putative sequences of total mass 17143 and 17171 Da, and the peptide YNVQYEGDNTFEPVY**M**TPE**F**JJFANQN**V**DR is shared with putative sequence of calculated mass 17255 Da.

Investigation into the heterogeneity of field vole glareosin has provided evidence for a number of point mutations contributing to a heterogeneous protein profile in most individuals. A mass discrepancy between intact masses 17138 and 17168 ± 1 Da is suggested to be a result of the mutation 91A>V (+28 Da), with considerable evidence of the mutation in two individual samples containing almost solely the peak 17168 ± 1 Da. Mutations are also proposed for the intact mass peaks 17236 and 17252 ± 1 Da. Considerable evidence for the peptides YNVQYEGDNTFEPVY[L>M]TPE[J>F]JJFANQN[A>V]DR and [G>D]Q[P>H]LT[T>P]EQ[Y>S]EK were observed, only together, which would account for the intact mass peak of 17236 ± 1 Da. Evidence for an additional mutation was found, EJTC[N>K]ESCDRJE, in all four individuals, which at 14 Da may account for a mass shift between 17236 and 17252 ± 1 Da. However, there was also considerable evidence for oxidation of the one or two methionine residues (see peptide maps) which would

explain this peak within a closer error margin. This is also the case for the peptide WVTVAJ[A>T]ADNVK, for which fragment ion spectra provided confirmation (see Figure 4.9). Whilst both of these mutations are seen with peptide level evidence, there is no confirmed correlation with observed masses from intact mass analysis. It is possible that additional heterogeneity is present, but at levels not corresponding to the major intact mass peaks. To confirm the sequence mutations observed with peptide-level evidence with more confidence, complete separation of proteins would be required, or alternatively, confirmation with genome data.

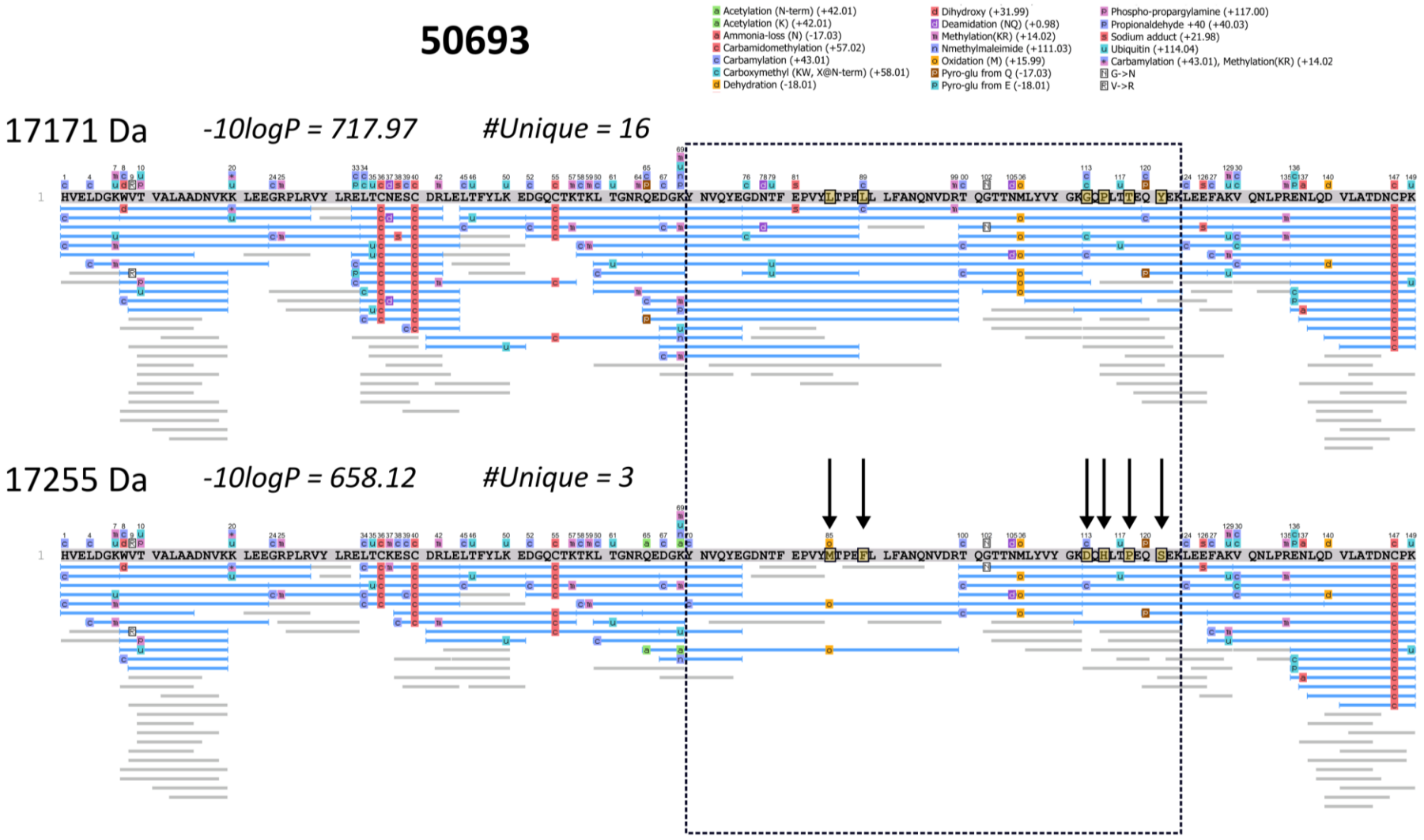
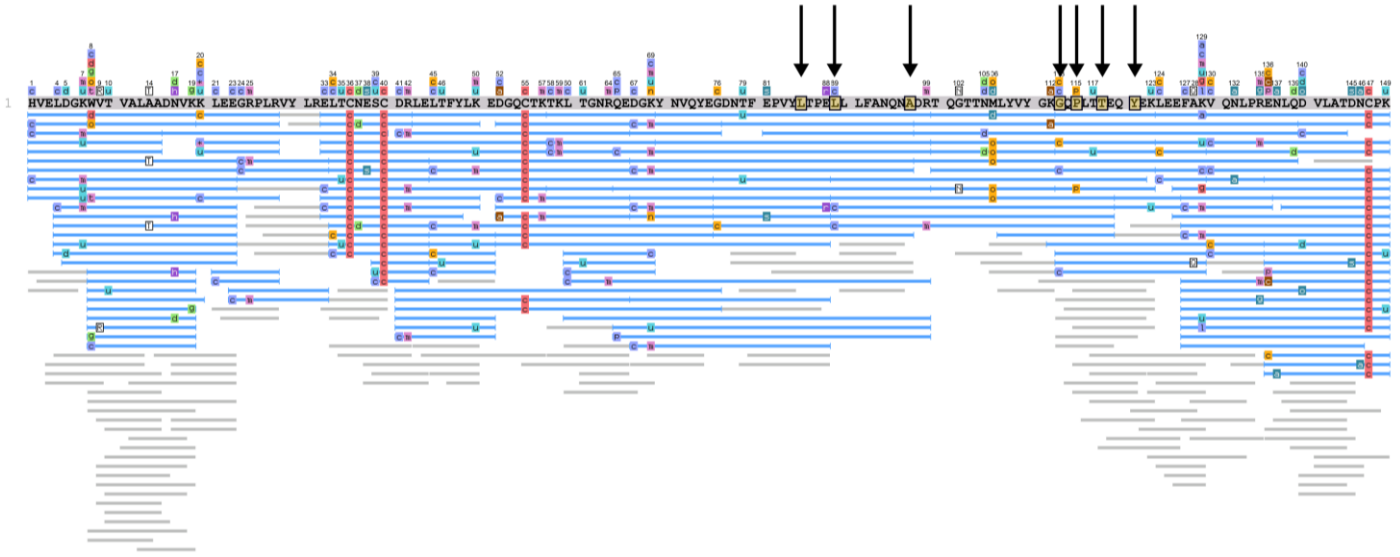


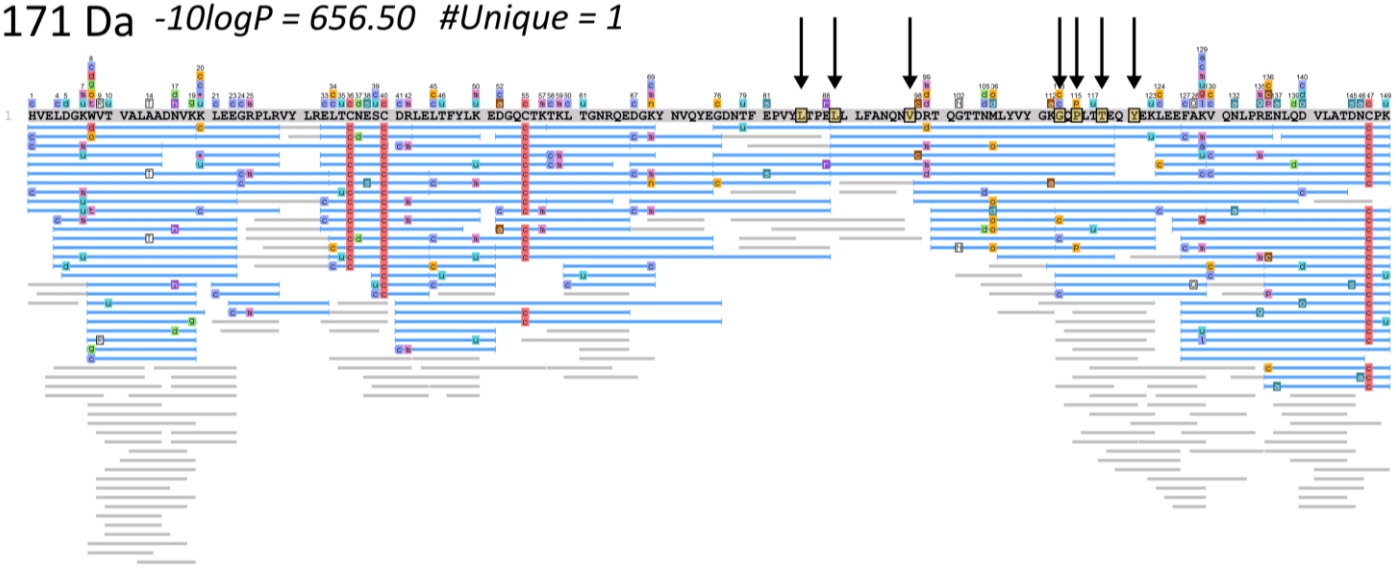
Figure 4.13 | Peptide map of field vole glareosin variants in a male sample with a protein profile dominated by a single mass peak, 17168 Da.
Peptides generated by proteolytic cleavage of field vole urinary protein from an individual sample dominated by a single mass peak, 17168 Da, were analysed by LC-MS/MS and inspected in PEAKS^{TM}. Peptide maps were generated for each hypothesised isoform.

50692

17143 Da *-10logP = 723.41 #Unique = 9*



17171 Da *-10logP = 656.50 #Unique = 1*



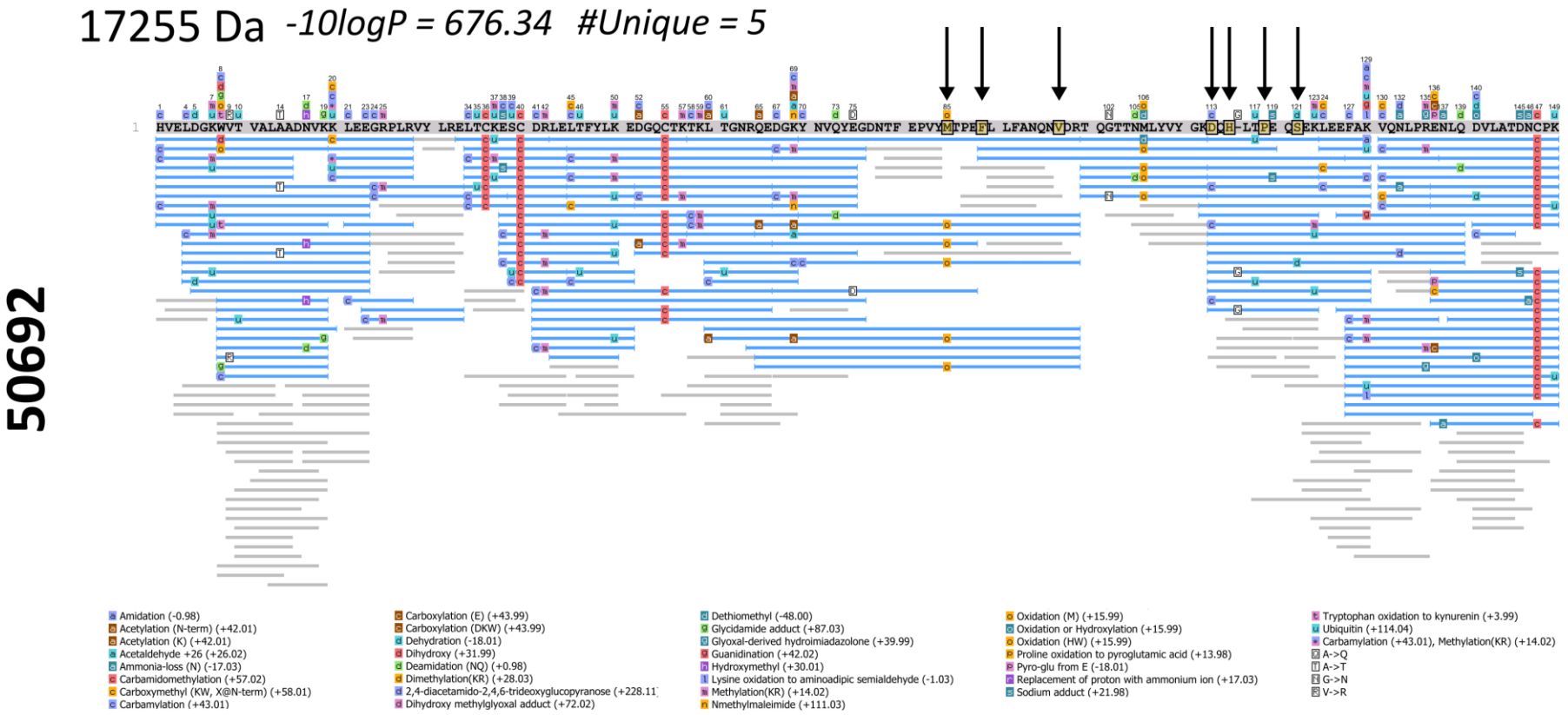


Figure 4.14 | Peptide maps of field vole glareosin variants in a male urine sample with four main peaks; 17138, 17168, 17236 and 17252 Da.
Peptides generated by proteolytic cleavage of field vole urinary protein from an individual with a more complex profile of four dominant mass peaks (17138, 17168, 17236 and 17252 Da), were analysed by LC-MS/MS and searched in PEAKS^{TM} SPIDER against a database comprising all *Rodentia* sequences in the UniProt database, bank vole glareosin, and four hypothesised sequences for field vole glareosin.

The proposed mutations in field vole glareosin may have functional relevance. Heterogeneity in the mouse major urinary proteins is critical to the role each proteoform plays, and can affect ligand binding capabilities. Each sequence variant, where every isobaric leucine or isoleucine residue was represented by leucine, was imported into the PHYRE2 server (Kelley *et al.*, 2015) to generate an approximate model based on a homologous structure. All sequences were modelled on the PDB structure 1DZK, porcine odorant binding protein from *Sus scrofa* complexed with pyrazine and were confidently modelled with the conserved structure of lipocalins, consisting of a beta-barrel of nine anti-parallel beta sheets. The resulting structures were viewed in PyMOL v2.0.7 (PyMOL, no date) and the proposed substitutions were located (Figure 4.15).

No mutations affected the internal calyx; all were outwardly-facing. The majority of mutations occurred at the 'closed' end of the lipocalin, according to standard lipocalin notation (Flower, 1996), with the exception of positions 97, which is located at the end of beta-sheet B at the open end of the lipocalin structure, and position 121, which is located on the alpha-helix. The proteoforms can be divided into pairs, within which only one amino acid is substituted; a 'light' pair of 17138 and 17168 Da, and a 'heavy' pair of 17236 and 17252 Da. It is substitutions occurring between pairs that contribute to the major changes at the closed end. Of note, the 'light' pair contains amino acids Leu85, Leu89, Gly113 and Pro115, that all contribute to an unusually hydrophobic external protein face, which in the 'heavy' pair are substituted to Met85, Phe89, Asp113 and His115, contributing to a more hydrophilic surface.

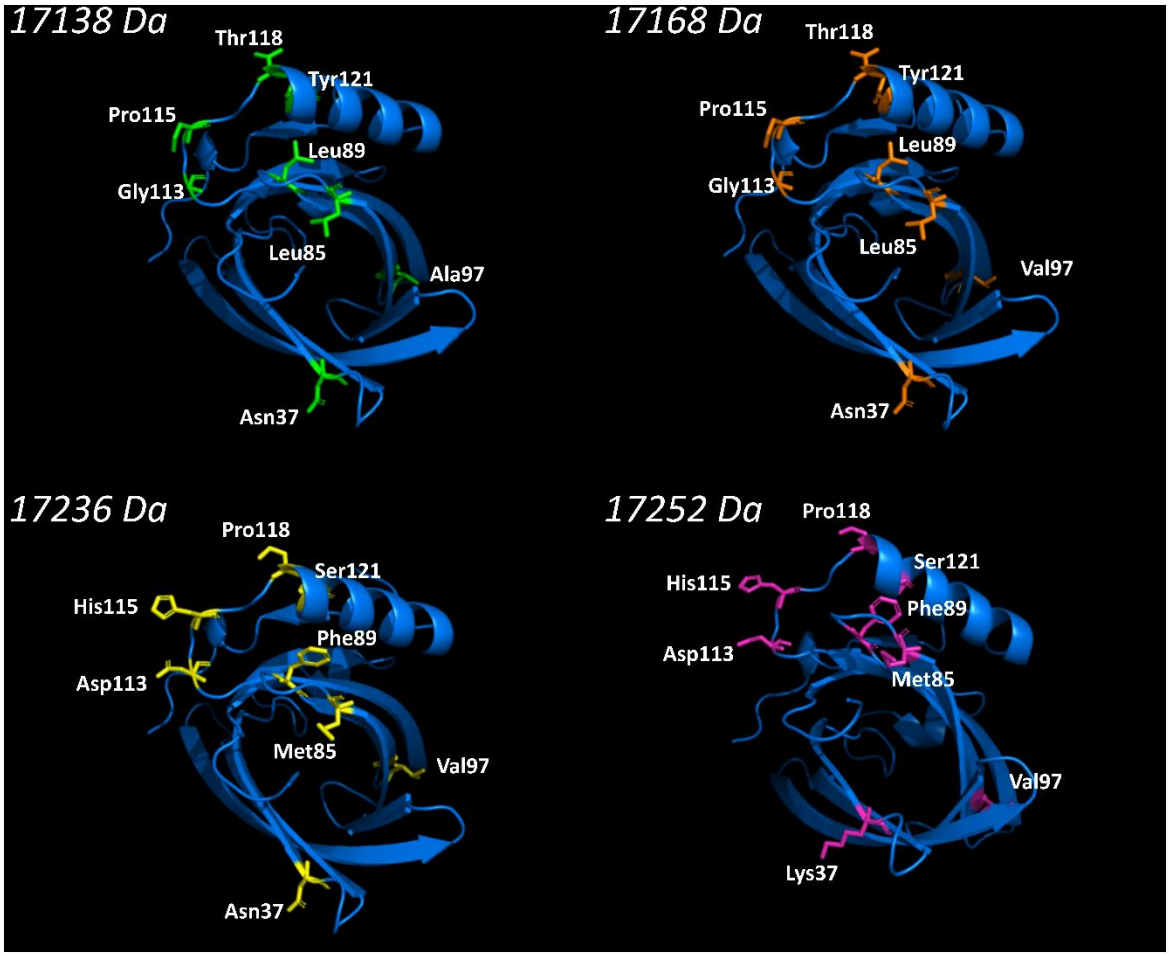


Figure 4.15 | Location and structure of proposed amino acid substitutions causing protein-level heterogeneity in field vole glareosin.
Four sequences containing amino acid substitutions, proposed to be the cause of protein heterogeneity observed in intact mass analysis of male field vole urine, were input into the PHYRE2 server (Kelley *et al.*, 2015) for structural homology modelling. The resulting structures, modelled on a porcine OBP structure, were analysed in PyMOL. Structures are denoted according to the associated intact mass and structure is viewed from the closed end of the lipocalin.

4.4.4 Identification of major urinary proteins in the female field vole

In contrast to bank voles, mature female field voles excrete notable amounts of protein into their urine. However in comparison to their male counterparts, these proteins not only run differently on SDS-PAGE (Figure 4.1) but upon inspection of intact mass profiles, mainly exhibit a cluster of peaks at approximately 21 kDa (Figure 4.3) rather than the 17 kDa masses explored previously.

To identify the main proteins, 2 µg protein from pooled urine of six female field voles was incubated with four different proteases, trypsin, glu-C, lys-C or asp-N, overnight. Resulting peptides were analysed separately by LC-MS/MS and data were explored in PEAKS™ by searching against a

database comprising all *Rodentia* sequences in UniProt, in addition to bank vole glareosin and the four potential field vole glareosin variants constructed as above.

No glareosin sequence was identified in pooled female field vole urine. A small amount of peptide level evidence for the odorant binding proteins (OBPs) was observed (Supplementary 4.6.3), but the main candidate to explain the 21 kDa mass observed was evidence of a major urinary protein (MUP), by identification of peptides homologous to MUP5, a truncated major urinary protein from *C.griseus* (UniProt Acession: G3HPK8), with a -10logP confidence score of 112.86 (Figure 4.16A). Sequence coverage was improved using a combination of the identified peptides, *de novo* tags and manual inspection of unidentified spectra, aligned against MUP sequences from *R.norvegicus*, *M.musculus*, *C.griseus* & *M.auratus* (Supplementary 4.6.4). A draft sequence, constructed from peptides sequenced *de novo* and amino acids from the hamster protein from the amino acid positions 41 to 57, was added to the previously modified database of UniProt *Rodentia* sequences containing field vole glareosin variants and data were re-searched. The draft MUP sequence was confidently identified (Figure 4.16B), however a portion of the protein remained unsequenced (Figure 4.16B, amino acids 41-58). Additionally, the total mass of the novel protein sequence, even with the unsequenced segment represented by a manually aligned portion of the *C.grisueus* protein, was 19138 Da, a mass deficit of over 2 kDa in comparison to the intact mass profile of 21 kDa. A possible explanation is the presence of an attached glycan, thereby masking the proteolytic sites nearby. This type of modification has previously been identified in other lipocalins including a mouse and a rat MUP (Cavaggioni and Mucignat-Caretta, 2000; Mechref *et al.*, 2000), in addition to aphrodisin and boar salivary lipocalin (Henzel *et al.*, 1988; Loebel *et al.*, 2000; Spinelli *et al.*, 2002). The N-terminal glycosylation site in aphrodisin and the site in SAL1 are both in the region of the missing peptide, and are aligned with a glycosylation motif in the *C.griseus* sequence. Glycosylation in the rat and mouse MUPs also occur close to this region (see Figure 1.4).

The novel sequences constructing the draft field vole MUP sequence were the among the highest abundance peptides sequenced in PEAKS^{TM}, and are therefore hypothesised to be contributing peptides to the 21 kDa mass peak observed in intact mass analysis. However, this requires further confirmation. Confirmation of the hypothesised glycosylation site, full sequencing *de novo* subsequent to removal of the attached glycan, and exploration of the heterogeneity of both the protein sequence and the carbohydrate are all avenues that should be taken before publication of this data, in a similar approach to that taken in chapter 5. Unfortunately, this work was beyond the scope of this thesis. Nevertheless, proteolytic digestion of protein from pooled female field vole urine provides strong evidence of a highly abundant protein, or proteins, related to the major urinary proteins, which is a vastly different protein landscape to the male equivalent.

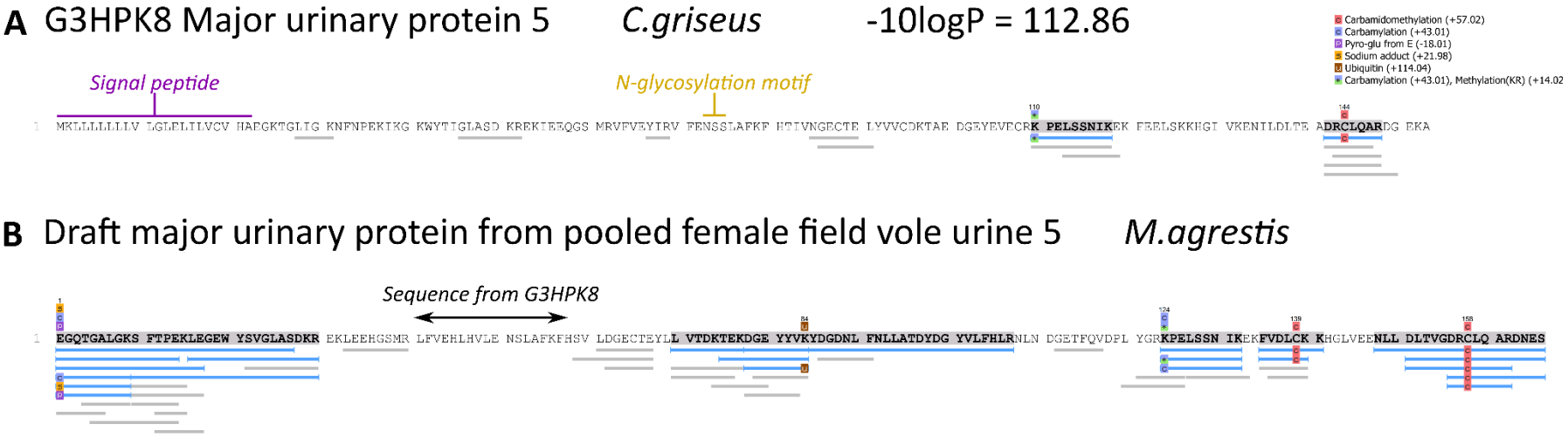


Figure 4.16 | Identification and sequence coverage of a major urinary protein in pooled female field vole urine.
Protein (2 µg) in from pooled female field vole urine was digested with four different proteases. Resulting peptides were analysed by LC-MS/MS and data were exported to PEAKS^{TM}. SPIDER searches identified a major urinary protein from *C.griseus*, or the golden hamster (A). Identified peptides were combined with *de novo* tags and unidentified spectra to improve sequence coverage (B).

4.4.5 Identification of field vole urinary proteins separated by PAGE and digested in-gel.

To assess the overall protein complement in the urine of mature field voles, 10 µL pooled urine from 13 males, and 10 µL pooled urine from six female field voles were analysed by SDS-PAGE and native PAGE (Figure 4.17). Distinct protein bands were excised and subject to proteolytic digestion with trypsin. Resulting peptides were analysed by LC-MS/MS and identified using PEAKS^{TM} SPIDER searches using a database of *Rodentia* sequences in UniProt, in addition to glareosin variants and the field vole MUP draft sequence. Candidates were selected based on the highest-scoring proteins, but took into account the total peak area of proteins, in addition to manually assessing unidentified spectra to ensure identifications were not missed due to lack of sequence similarity from cross-species matching.

Unsurprisingly, the slowest migrating protein band in both males and females, and in both native and SDS-PAGE was uromodulin, a common urinary tract protein of approximately 70 kDa. Similarly, the second slowest migrating band was albumin, although a corresponding band was only faintly seen in pooled female urine by SDS-PAGE analysis, so was not subjected to in-gel digestion. Peptides from the protein band resolving at approximately 25 kDa in the SDS-PAGE gel in both males and females identified the band as field vole major urinary protein(s). This is indicative that male field voles also express this protein, or proteins, but this expression is secondary to that of field vole glareosin. MUP peptides were only identified in a single band in female urine when analysed by native PAGE, although 3 MUP peptides were identified in the lowest native PAGE gel band from male urine. However, the protein band was predominantly a variant of field vole glareosin. Field vole glareosin was only identified in males, from both SDS and native PAGE analyses. In both cases, the two fastest-resolving protein bands were identified as glareosin. Whilst the glareosin variants with total sequence masses of 17143, 17171 and 17241 Da were all identified in each band, the upper band in both cases was attributed to pair ‘A’ (17143 and 17171 Da, which only differ by one amino acid) and the lower to type ‘B’ (17241 Da), in which the central region contains 6 mutations that differ to type ‘A’. As the sequences are so similar, this relied on comparison of individual peptides containing the mutations, and identification of protein bands was based on the number and scoring of unique peptides from ach variant identified. Pooled female urine contained one other protein band (resolving at approximately 18 kDa) that was most closely identified as OBP2 from *M.glareolus*. However, the low protein content of this protein band reduced the amount of peptide information available. Nevertheless, OBP2 peptides were also seen in the fastest-resolving bands in both males and females from both native and SDS-PAGE analysis, providing considerable evidence of an OBP2-like protein present in both male and female urine at a low level.

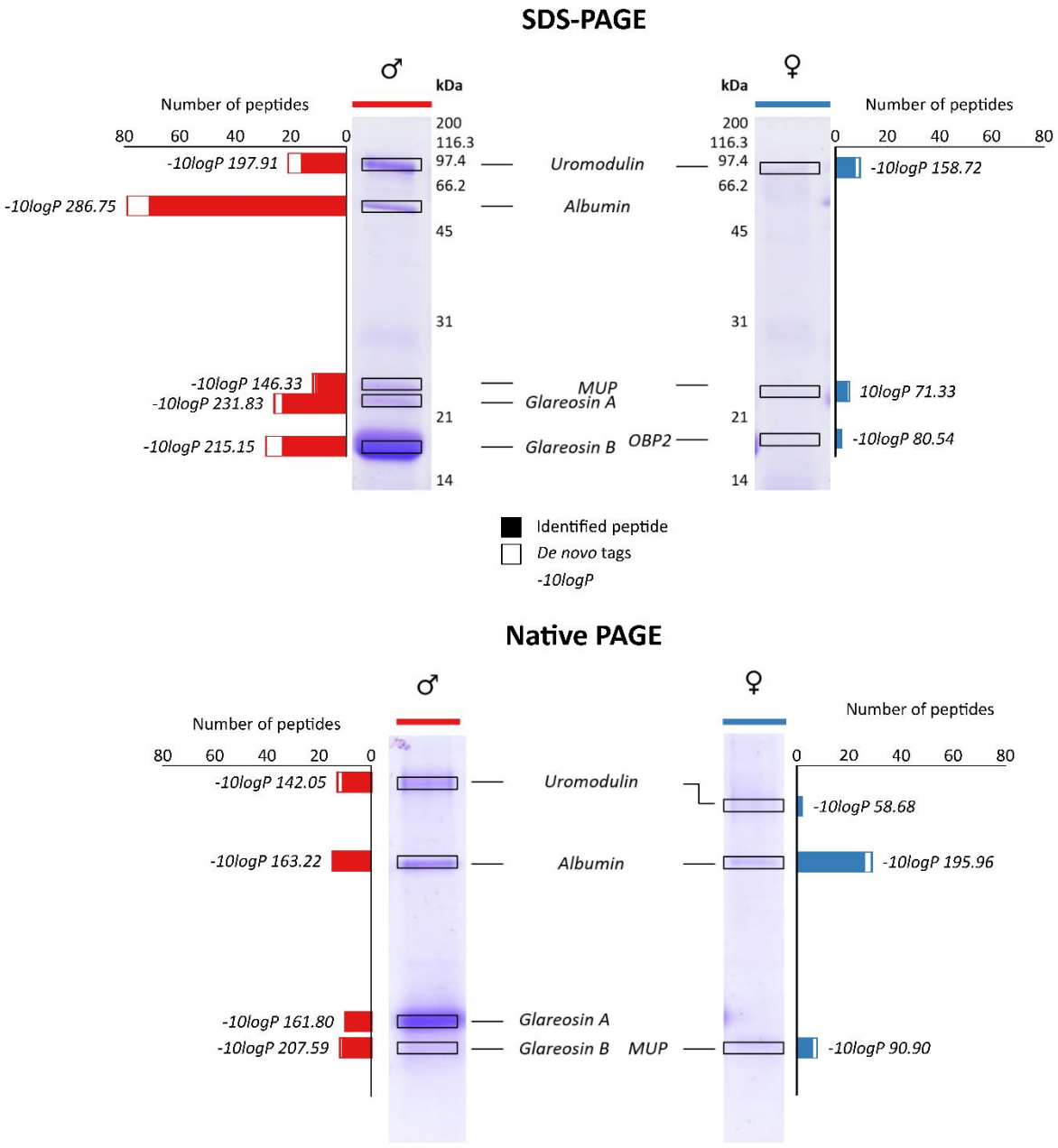


Figure 4.17 | In-gel identification of protein bands from SDS (top) and native (bottom) PAGE analysis of urine pooled from mature male and female field voles.

Pooled urine from mature male and female field voles (10 µL) was analysed by 15% SDS and native PAGE. Visible protein bands stained by coomassie were excised, and encompassed proteins were subject to proteolytic digestion with trypsin. Peptides were analysed by LC-MS/MS. Data were searched against a database of *Rodentia* sequences, including bank vole and field vole glareosin sequences, in addition to the novel field vole MUP sequence. Top-scoring, highly abundant proteins identified were considered as candidates, and the number of supporting peptides identified, in addition to the number of *de novo* tags, were recorded to support identification.

4.4.6 Global proteomics of mature field vole urine

Overall complexity of mature field vole urine has been explored by PAGE and intact mass analysis. Major proteins in male and female urine have been identified and sequenced to some degree, and in-gel digestion has confirmed these sequences to be representative of the predominant urinary proteins. However, additional layers of complexity can be present at lower levels. In the bank vole, additional OBP-like sequences were identified that suggested heterogeneity in the OBP proteins present in bank vole urine, aside from the dominating protein of glareosin.

To investigate the lower levels of complexity in field voles, six female and six male urine samples were normalised to creatinine, to correct for urine dilution, and digested with trypsin. Resulting peptides were analysed by LC-MS/MS and analysed in PEAKS^{TM}. Following the assessment of using large, multispecies databases for quantification versus single organism databases (see Chapter 3), label-free quantification was performed on identifications from a database of all house mouse sequences in UniProt, in addition to novel protein sequences. These included field vole glareosin, the field vole MUP draft sequence and an OBP3-like sequence generated from analysis of juvenile field vole samples (see section 4.4.7.2). Peptide filtering was results-based (also discussed in Chapter 3), and peptide quality was restricted to ≥2. Average peak area was set to those ≥1e^5. A protein significance value was calculated using the PEAKSQ method, but was not restricted, and neither was the fold change. The number of unique peptides was left at 1, due to the sequence similarity between glareosin sequences but is worth considering. After trypsin and other contaminants were removed from the analyses, 29 quantifiable proteins remained.

The novel field vole MUP sequence was the most abundant protein identified on average for both males and females. Comparison of male and female samples revealed seven proteins with a significance of over 25 and a fold change of over 2. Glareosin (17171 Da) and field vole OBP3 were both quantified at far higher values in males compared to females (Figure 4.18). Field vole MUP also had a significant fold increase in male samples, albeit to a lesser degree than glareosin or OBP3. However, albumin, uromodulin, serotransferrin and transthyretin are all common urinary proteins which should be comparable in males and females, and an increase of these proteins suggests that a bias may have been introduced when normalising to creatinine.

Additionally, quantification of field vole glareosin proved problematic. The main differences between the field vole glareosin variants lie within a central portion of the protein that is not easily accessible with a trypsin restricted digest. For future quantification of the protein, an alternative proteolytic mechanism should be considered. As a result, one variant of type ‘A’ glareosin (17143

Da) was quantifiable in three male samples, and the other type 'A' variant quantifiable in the other three.

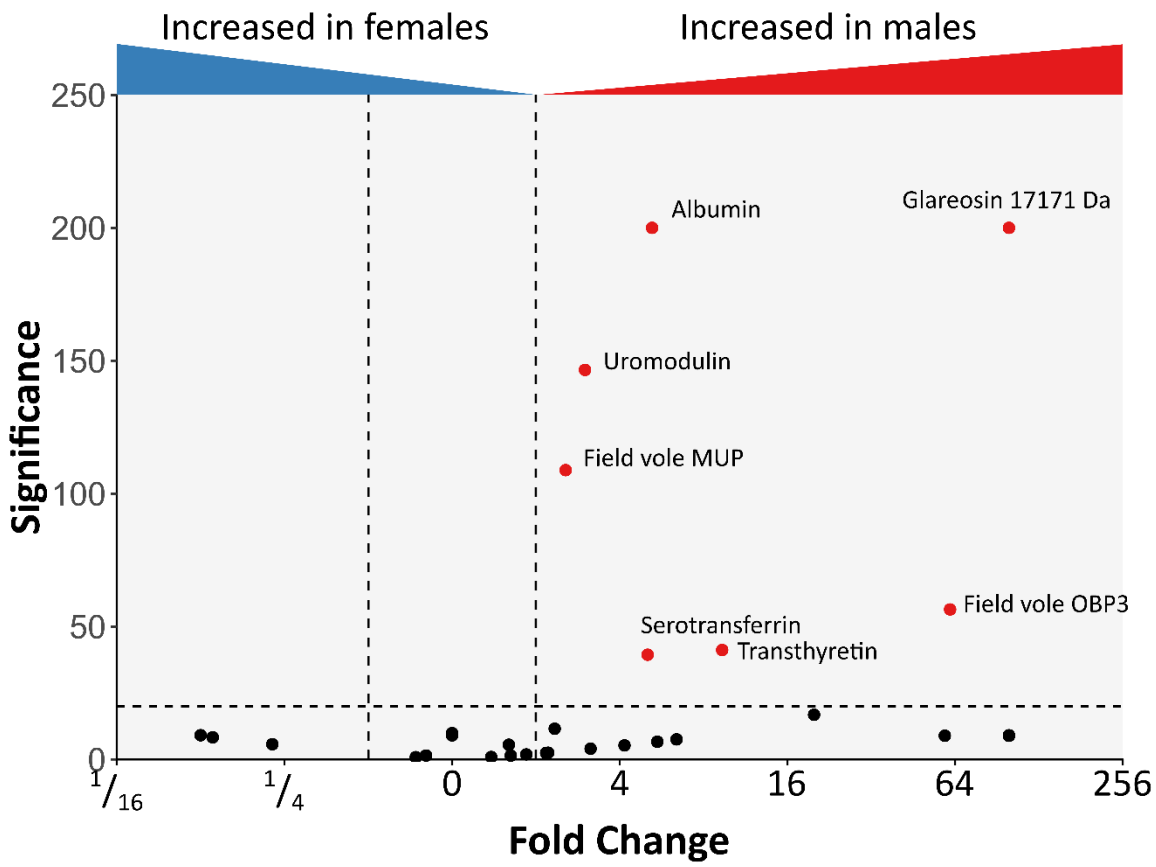


Figure 4.18 | Label-free quantification of urinary proteins in mature field voles.
Urinary proteins from mature male (n = 6) and female (n = 6) field voles were digested with trypsin and subject to LC-MS/MS analysis. Protein identification and subsequent label-free quantification was performed in PEAKS™, by searching against a database comprising all *Rodentia* sequences in UniProt, plus bank vole and field vole glareosin, field vole MUP draft and field vole OBP3 draft (see section 4.4.7.2). Proteins with a significance over 25 and a fold change over 2 are indicated.

4.4.7 Urine from sexually immature field voles

4.4.7.1 Initial assessment of urinary protein content of immature field voles

Whilst scent signalling is usually associated with sexual selection and mating behaviour, juveniles can also utilise olfactory cues, for example juvenile mice secrete the exocrine-gland secreting peptide 22 (ESP22), a lacrimal gland peptide which supresses sexual response from mature mice (Ferrero *et al.*, 2013). A small scale investigation was initiated to assess the urinary protein output of juvenile field voles.

Overall urinary protein output of male (n=5) and female (n=4) juvenile individuals was assessed using a Bradford assay. Sex did not significantly affect total output of protein ($\chi^2(-1) = 2.54$, $p = 0.15$) or creatinine ($\chi^2(-1) = 1.88$, $p = 0.21$), and protein profiles normalised to 0.3 µg creatinine, as inspected by SDS and native PAGE, were variable (Figure 4.19). Three distinct bands were observed within the high molecular weight region (>45 kDa), but within the lower molecular weight region of SDS-PAGE, multiple protein bands were observed at approximately 21 kDa and 16 kDa. Native PAGE of the same samples separates three consistent bands likely to correspond to the higher molecular weight bands in SDS-PAGE, and two other distinct bands resolving faster in the gel, where in each individual, one is usually far more strongly stained than the other. No statistical comparison was performed with mature field vole urine samples due to the low number of juvenile samples, but male and female juvenile urine had a lower average protein content of 667 µg/mL and 365 µg/mL compared to 1124 µg/mL and 663 µg/mL in mature male and female field vole urine samples, respectively. Protein profiling by PAGE was also less consistent than for adult samples, where multiple protein bands of varying intensity were observed in the lower molecular weight region of the gel, in contrast to the strong consistent doublet of field vole glareosin observed in mature males, and the fainter 21 kDa band observed in mature females.

Urinary protein (1 pmol) was then analysed by ESI-MS to assess protein profiles. Examples are displayed in Figure 4.20. Spectra were heterogeneous and variable across individuals, however, a region of distinct intensities between 16.5 and 17.5 kDa was consistent in all individuals. Masses within this region were variable, but in all males and three of four females the peak 16893 ± 1 Da was observed, and was the most intense peak in one male and two females. Masses corresponding to field vole glareosin variants were also observed in one male juvenile. Another region of peaks was detected at approximately 21 kDa in three males and one female. There was high variability between samples, both inter-sex and between individuals, and no definitive patterns were identified. Overall sample complexity was further investigated using a bottom-up proteomics approach of whole urine samples.

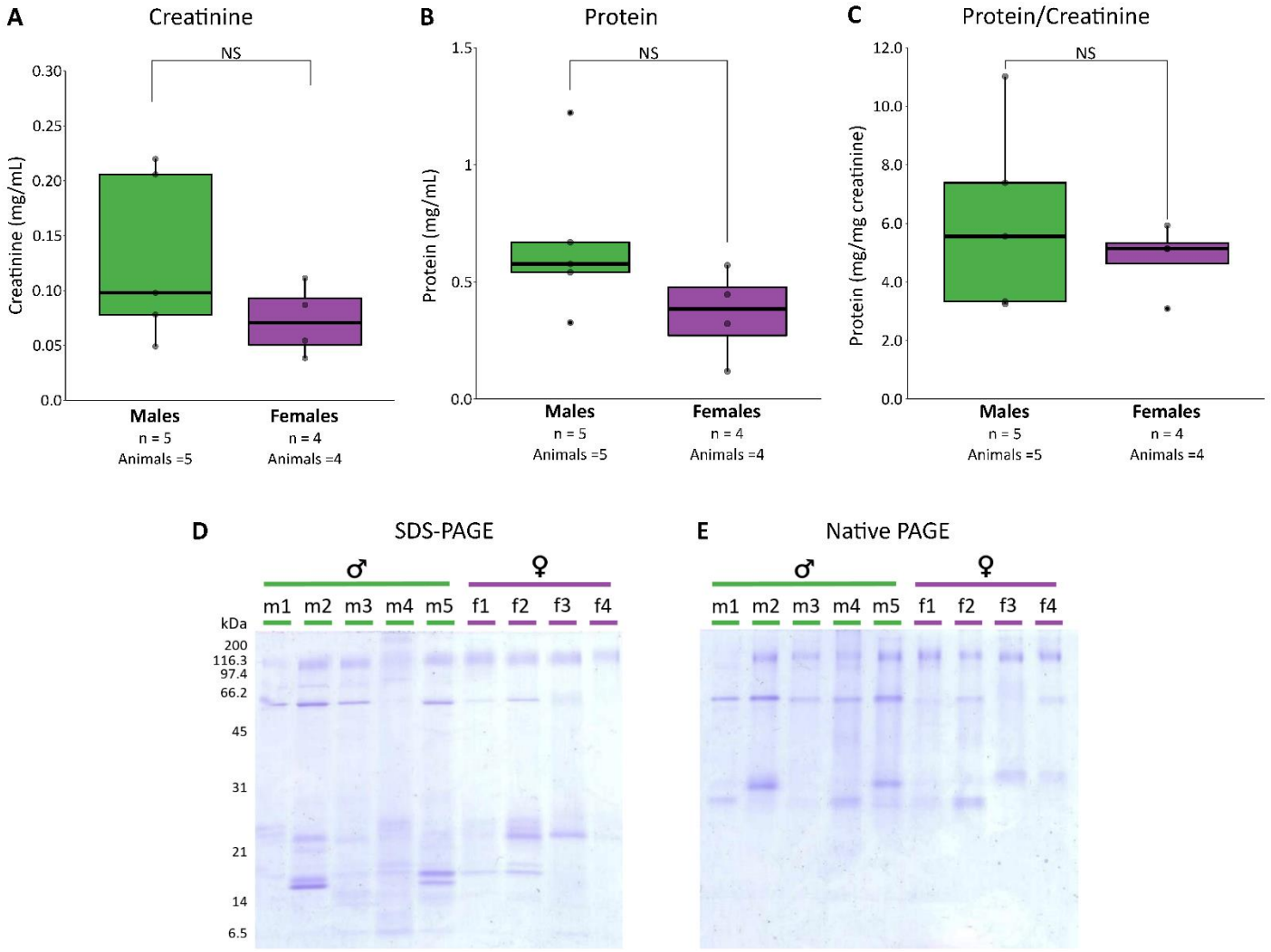


Figure 4.19 | Initial assessment of protein output in the urine of juvenile field voles.
Male (n=5) and female (n=4) juvenile field vole urine samples were assessed for creatinine output, as a measure of urinary dilution (A), protein output (B) and protein relative to creatinine (C). Overall protein profiles (normalised to 0.3 µg creatinine) were initially assessed by SDS-PAGE (D) and native PAGE (E).

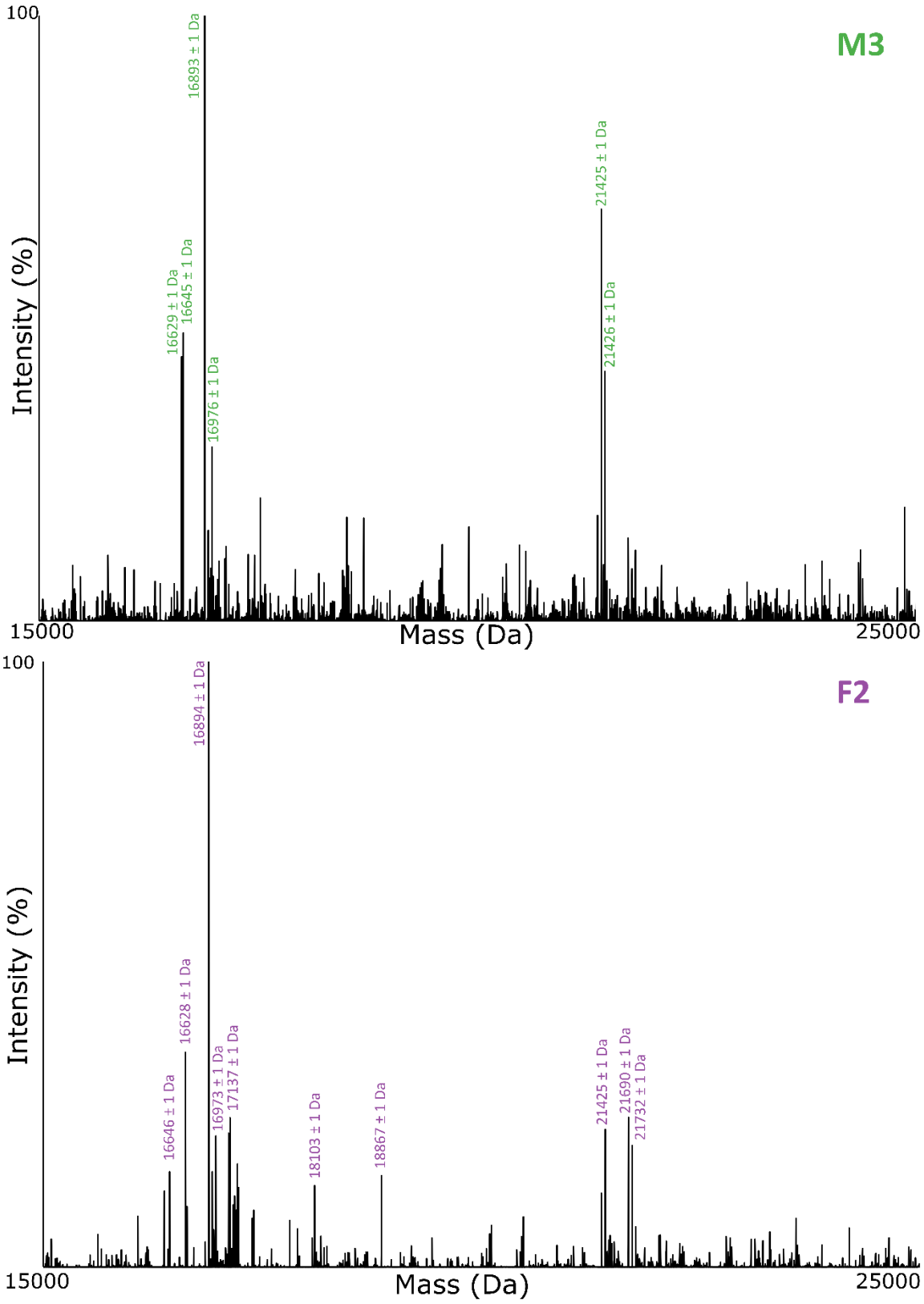


Figure 4.20 | Intact mass analysis of urinary proteins from juvenile field voles.
Proteins in urine from juvenile field voles were analysed by ESI-MS. After initial deconvolution across a wide mass range (5000 – 100000 Da), main peaks were located within the lower molecular weight region and spectra were deconvoluted between 5000 and 30000 Da with a smaller mass tolerance.

4.4.7.2 Discovery proteomics of urine from juvenile field voles

Protein (5 µg) in urine samples from 5 male and 4 female juveniles were digested with trypsin, analysed by LC-MS/MS and identifications were made by searching a database of *Rodentia* sequences with additions made of bank vole glareosin, field vole glareosin (all variants), and the field vole MUP draft sequence using the PEAKS^{TM} SPIDER tool. A false discovery rate (FDR) of 1% was applied and label-free quantification was applied to high quality identifications. Quantified proteins were filtered as follows. Peptides with a peak area lower than 1 x 10^5, and a quality score less than 2 were excluded, and proteins were filtered out if they were identified by only one unique peptide. Contaminants and trypsin were also removed from the analysis. Data from the remaining 24 proteins were subject to hierarchical clustering (Figure 4.21), to indicate how similar individual samples are based on the abundances of identified proteins.

Samples were not clustered by sex, indicating that the protein abundance profiles of juvenile field vole urine are not different enough to be grouped by sex, or there is not enough information in this investigation to observe differences. The highest abundant protein in all females and three males was the draft MUP sequence. This most likely accounts for the 21 kDa peaks observed from intact mass analysis, given the similarity to mature female profiles. In the other two males, glareosin (variant 1, total sequence mass 17143 Da) and OBP3 (*Myodes glareolus*) were the most abundant proteins.

Glareosin was not quantifiable in any females, but was in four males. The male sample in which glareosin was the highest abundant protein had a protein profile dominated by the peaks 17138 Da and 17168 Da, similar to that of mature males. Another male sample also had notably high levels of glareosin, however in this sample OBP3 was identified as the most abundant protein.

Of note, OBP3 was identified through cross-species matching at high levels in all juvenile field vole samples. It is also worth noting that one of the few common peaks observed in almost all protein profiles had a mass 16893 ± 1 Da, and given the strong evidence for OBP3 in all juvenile samples, this is a candidate mass for the source of the peptides homologous to bank vole OBP3.

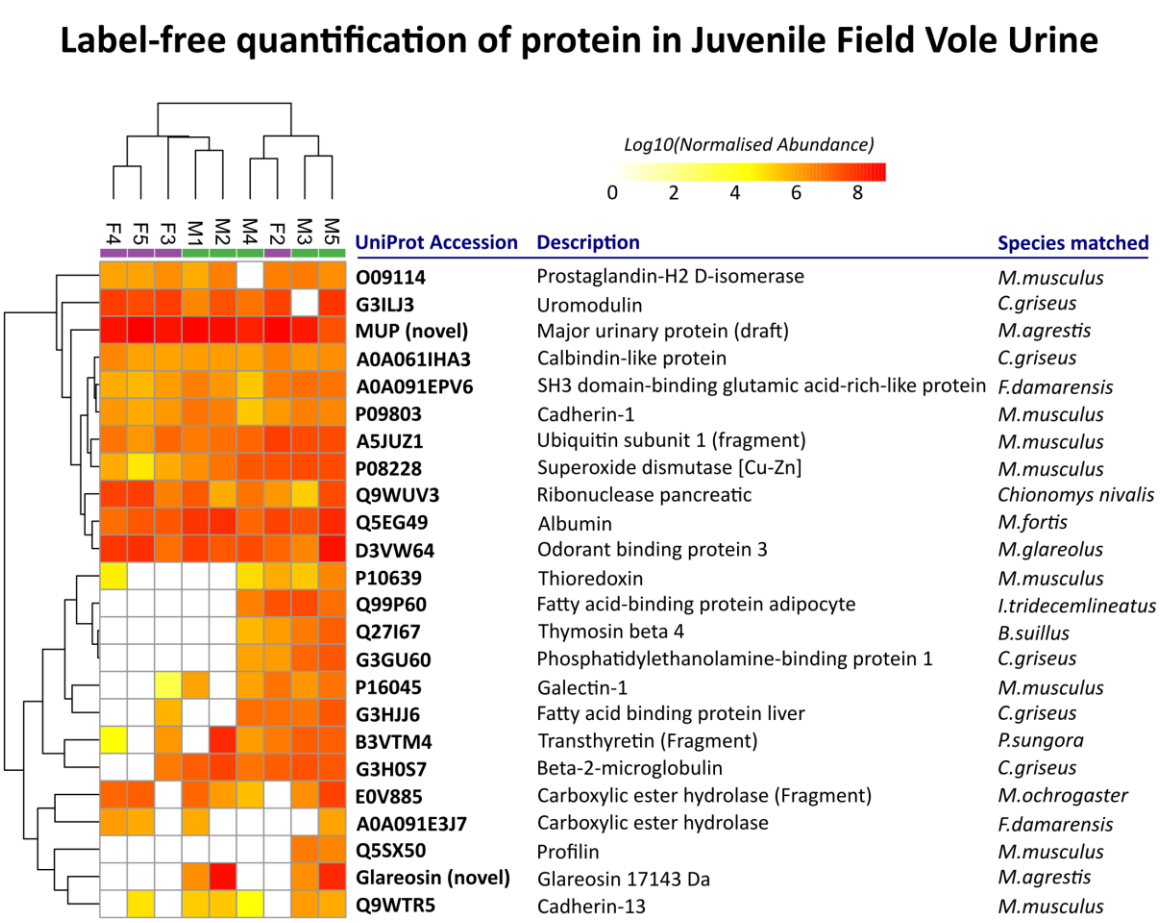


Figure 4.21 | Label-free quantification of proteins detected in juvenile field vole samples. Peptides (5 µg) from tryptic digests of urinary protein from juvenile male and female field voles were analysed by LC-MS/MS and analysed in PEAKS^{TM}. Proteins confidently identified by more than 1 unique peptide were quantified and subject to hierarchical clustering.

To investigate the 16893 ± Da peak further, which is also present at low levels in mature field vole urine, protein in two samples from juvenile female field voles were digested with four different proteases (trypsin, glu-C, lys-C or asp-N). Protein profiles of these two samples exhibited particularly high abundance of this mass peak. Parallel digests were analysed by LC-MS/MS and examined in PEAKS^{TM}. Inspection of suggested mutations and confidently sequenced *de novo* tags of bank vole OBP3 resulted in the construction of a draft field vole OBP3 sequence, with a mass of 16986 Da (Figure 4.22). Unfortunately there were some areas of the sequence that were less confidently elucidated, and full sequencing *de novo* was not possible. Intact mass data displayed a heterogeneous profile, and despite the abundance of 16893 ± 1 Da, surrounding peaks may stem from heterogeneity of the OBP-like protein, as in bank voles, and other homologous peptides are likely to prevent confident sequencing. However allowing for reduction of 2 hypothetical disulfide bonds, a conserved structural feature of OBP lipocalins, this mass would reduce the observed intact

mass to 16982 Da, a mere 89 Da difference to the consistently observed mass of 16893 ± 1 Da. The draft sequence was inserted into the database and the discovery proteomics data were re-searched, providing strong peptide-level evidence (Figure 4.23). Juvenile samples could therefore provide an ideal sample base from which to further analyse field vole 'OBP3', and potential heterogeneity in field vole odorant binding proteins not deciphered in this preliminary investigation. Whilst not completed, this sequence was used to represent a proposed field vole OBP-like protein in subsequent identifications.

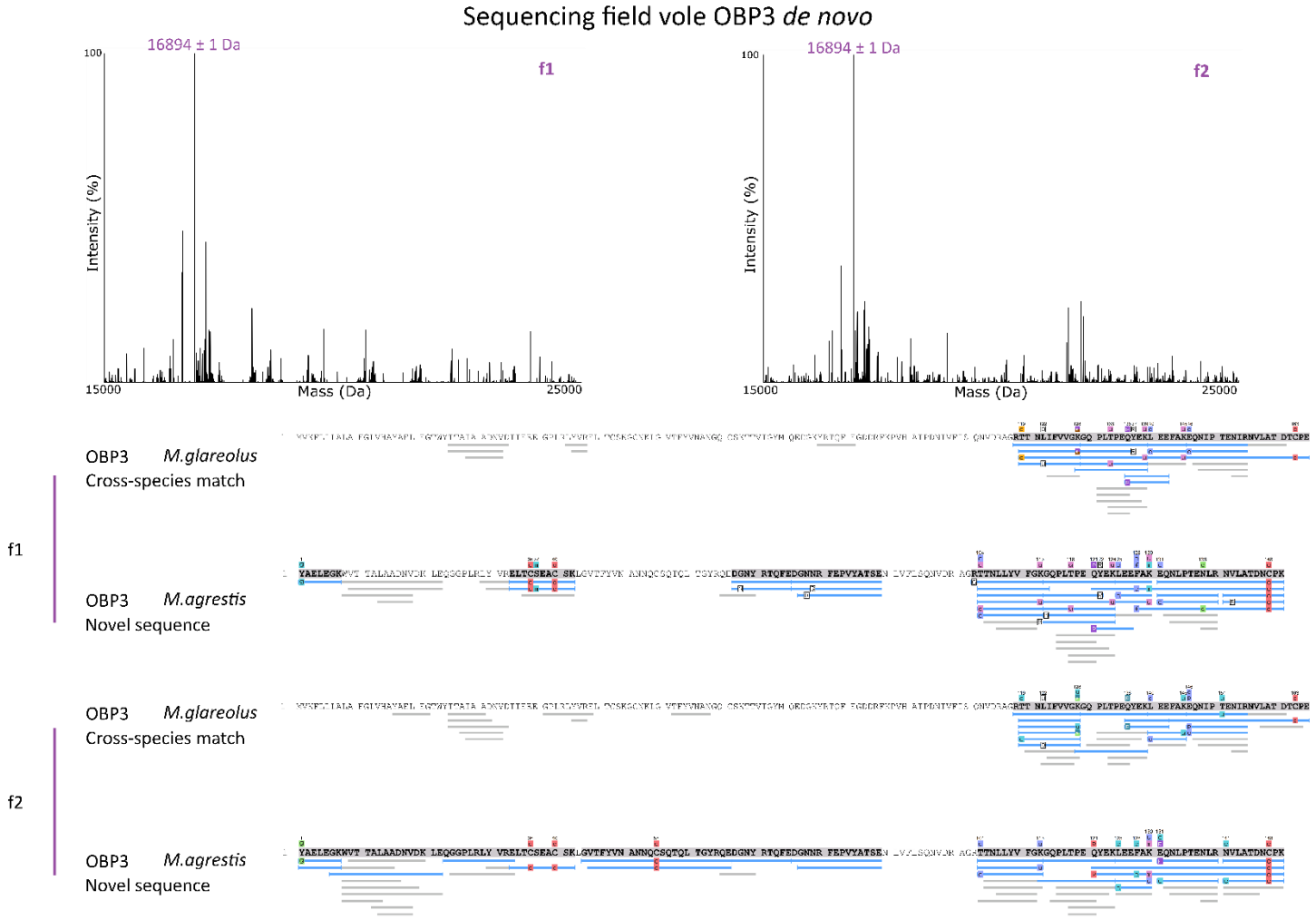


Figure 4.22 | Sequencing a draft of field vole OBP3 from multiple protease digestion of two juvenile female field vole urine samples.
Two urine samples from juvenile female field voles (f1 and f2; lanes 6 and 7, respectively, from Figure 4.19D&E) were selected to investigate the 16893 ± 1 Da mass peak observed from intact mass analysis. Each were digested in parallel reactions with trypsin, glu-C, lys-C and asp-N and resulting peptides were analysed by LC-MS/MS. Cross-species matching identified bank vole OBP3, which was used as a scaffold to construct a draft field vole sequence.

Initial sequencing of field vole OBP3 from juvenile urinary samples

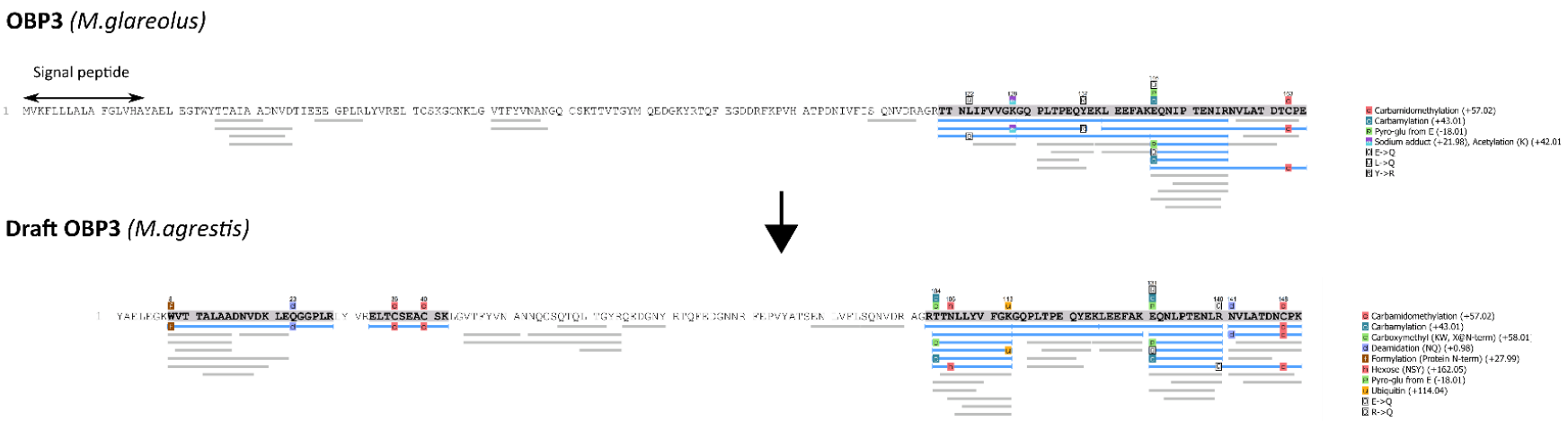


Figure 4.23 | Initial sequencing of field vole OBP3 from analysis of juvenile field vole urine.
Peptides (5 µg) from tryptic digests of urinary protein from juvenile male and female field voles were analysed by LC-MS/MS and analysed in PEAKS^{TM}. OBP3 from *M.glareolus* was confidently identified and a draft field vole OBP3 was constructed from suggested mutations and confident *de novo* tags.

4.4.8 *A preliminary exploration of scent mark protein content in the field vole, Microtus agrestis*

4.4.8.1 *Initial assessment of protein output in field vole scent marks*

Whilst urine can be an important vehicle in the conveyance of semiochemical information, in-house observations by the Mammalian Behaviour & Evolution Group at the University of Liverpool suggest there is a notable difference between urination and scent marking behaviours in field voles, and both have potential to play important but distinct roles in the transmission of social cues. Lipocalins, at varying degrees of heterogeneity, are expressed to considerable levels in mature and juvenile field vole urine. A preliminary assessment of the protein output in field vole scent marks was taken to investigate the overall level of complexity of scent marks compared to urine, and to identify putative chemosignalling proteins.

Scent marks were obtained from 10 mature males and 6 mature females whilst housed in their home cages. Deposits were swabbed with cotton buds soaked in 50 µL PBS and liquid collected by centrifugation. The overall protein content was assessed by Bradford assay and complexity was analysed by SDS-PAGE (7.5 µL) (Figure 4.24). Sample collection and SDS-PAGE analysis were undertaken by Holly Coombes (Mammalian Behaviour & Evolution Group, University of Liverpool).

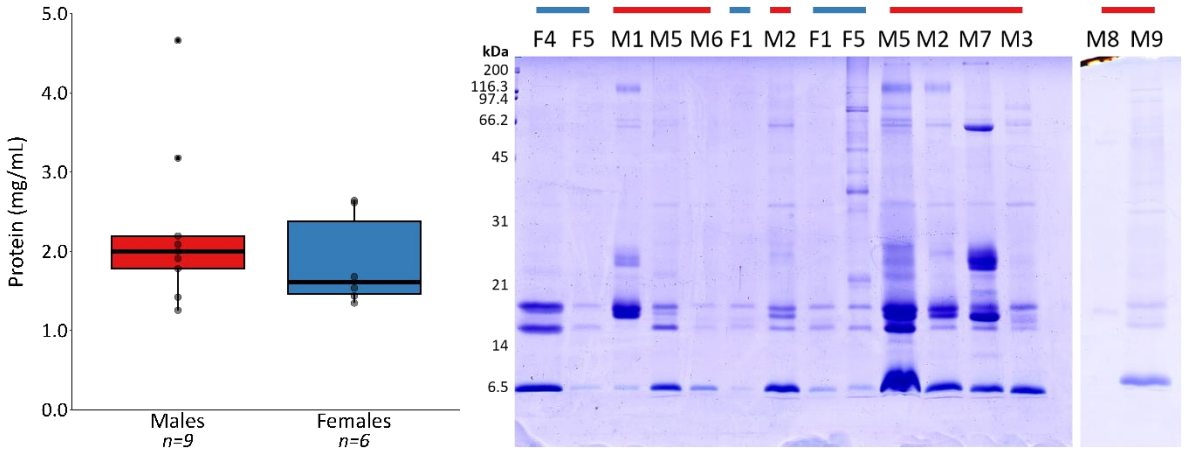


Figure 4.24 | Initial assessment of the protein content of field vole scent marks.
Scent marks deposited by mature male and female field voles were collected on swabs wetted with 50 µL PBS. Protein content was assessed by measuring concentration (A) and by SDS-PAGE* analysis of 7.5 µL recovered sample (B). *SDS-PAGE was performed by Holly Coombes, image used with permission.

Sex did not affect protein output of scent marks ($\chi^2(-1)=0.7104$, $p=0.4146$), although female protein output was lower by 399.9 ± 474.4 µg/mL (mean ± SE). However, higher sample numbers could further investigate significant sex differences. SDS-PAGE analysis revealed three strong, consistent protein bands in female scent marks, at approximately 19 kDa, 16 kDa and 6.5 kDa. The same protein bands were seen in male scent marks, but an additional band was also seen at approximately 18 kDa. Additional protein bands were also seen at higher molecular weights of approximately 25 and 35 kDa.

Protein profiles of scent marks were investigated by intact mass analysis. Unfortunately, sampling techniques resulted in interference of the ion signal and m/z charge envelopes were not generated, with the exception of two samples, one male (M5; lane 10 Figure 4.24B) and one female (F4; lane 1 Figure 4.24B). Both deconvoluted spectra shared three main peak intensities; 16628, 16648 and 16788 ± 1 Da (Figure 4.25).

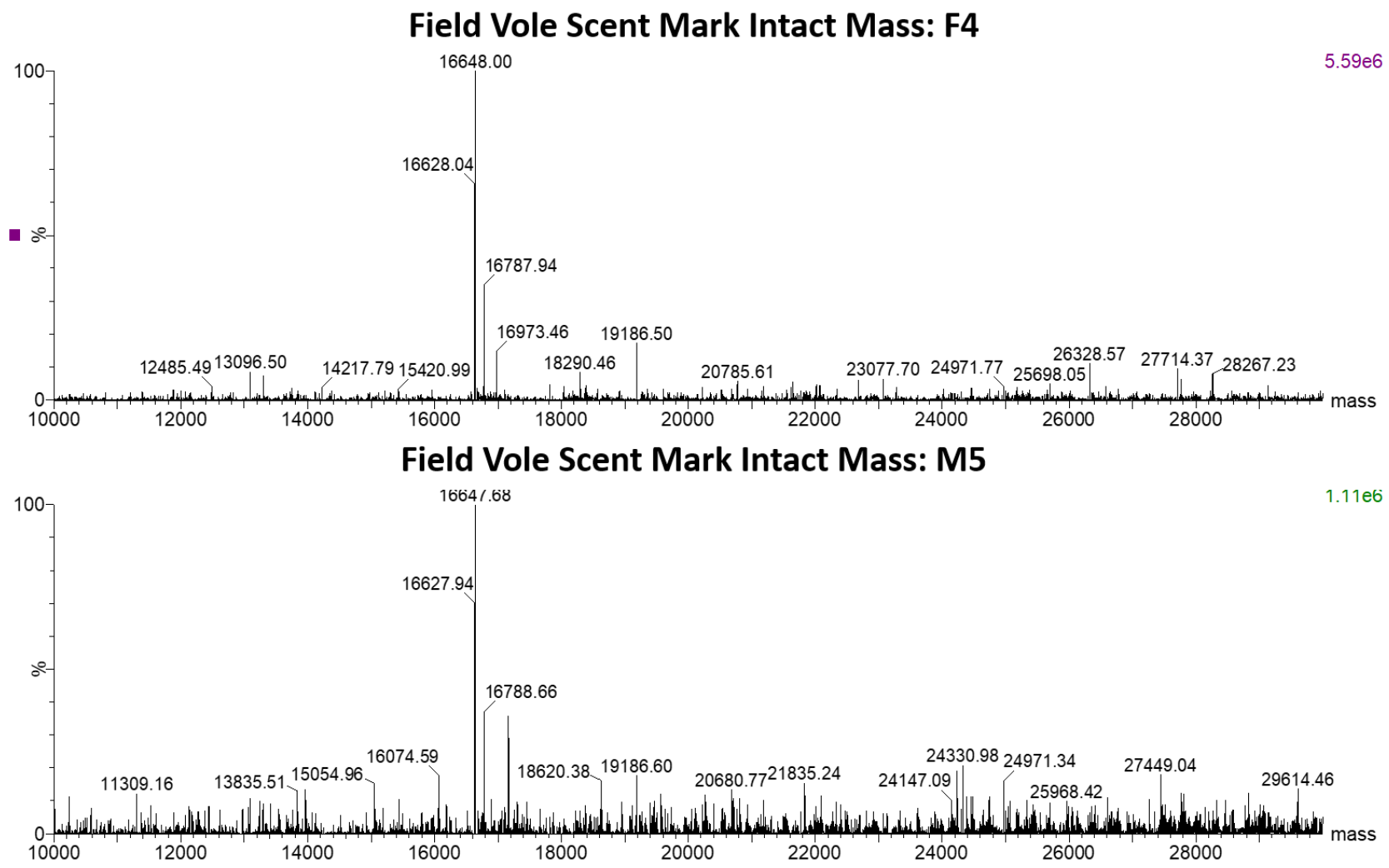


Figure 4.25 | ***Intact mass analysis of field vole scent marks.***
Protein (4 pmol) from scent marks of male and female field voles was analysed intact by ESI-MS. Sampling interference prevented spectral deconvolution of most samples, with the exception of one male and one female sample (above)

4.4.8.2 *Identification of field vole scent mark proteins separated by SDS-PAGE*

To review the types of protein present in field vole scent marks, individual protein bands from SDS-PAGE analysis of scent marks from mature male and female members of the species were excised and subject to proteolytic digestion. Abundant, consistent bands were the focus, to identify consistent patterns, and most protein bands to be analysed resolved to a molecular weight lower than 31 kDa. Replicates from three individuals for each sex, where possible, were taken and proteins identified from a combination of protein score, peak area, spectral counts and manual investigation of the unidentified spectra (Figure 4.27Figure 4.26 & Figure 4.27).

Abundant and consistent protein bands at three resolved molecular weights in female scent marks were identified as OBP3-like (approximately 18 and 16 kDa) and a member of the secretoglobin family (approximately 6.5 kDa). The OBP3-like protein bands were identified by homology to bank vole OBP3. Other proteins were also identified in some bands, particularly the highest molecular weight band analysed (18 kDa), including OBP2 & OBP1 (*M.glareolus*) and field vole glareosin, but in all cases OBP3 was the most confidently identified. Secretoglobin was identified by peptides homologous to *C.griseus* proteins with UniProt accessions G3IM86 (prostatic steroid-binding protein C1) and G3IM12 (secretoglobin family 1D member 2), both members of the secretoglobin family. It was the most abundant and most confidently identified protein from digestion of the band resolving at approximately 6 kDa for F4 and F1. Digestion of the equivalent band in F5 revealed peptides that most confidently identified field vole glareosin and bank vole OBPs. Prostatic steroid-binding protein, or secretoglobin, was also identified as a highly abundant protein, and more confident identification of lipocalins was thought likely to be due to close proximity on the gel to a highly abundant band in males (see Figure 4.24, lane 10).

Identification of male proteins was considerably more difficult, and protein identification was largely inconclusive (Figure 4.27). The fastest-resolving protein band, like in female scent marks, was identified as secretoglobin in one individual (M5, -10logP = 141.86, to *C.griseus* UniProt accessions G3IM86 and G3IM12). The equivalent band in remaining individuals was also identified as secretoglobin, however, this was not by cross-species matching, but manual inspection of high abundance *de novo* peptides with good ALC (%) scores. The comparatively small number of identified proteins from in-gel digestion of these bands were not derived from PSMs of high enough abundance to provide conclusive identification, in comparison with the abundance of the *de novo* sequences that remained unidentified and therefore suggest missing identification. The same situation occurred in identification of the bands at approximately 25 kDa, in which case no identification could be made, even by inspection of peptides sequenced by the software *de novo*. A similar problem persisted with the excised protein bands resolving between 16 and 20 kDa. For two

individuals (M5 & M2), two protein bands (approximately 17 & 18 kDa) revealed protein identifications of field vole glareosin and bank vole OBP3, however discrepancies meant that no consistent identifications were made. No identification was made for the equivalent bands in M6, although manual inspection of the unmatched *de novo* sequences revealed abundant glareosin- and OBP-like peptides. For example, analysis of the 17 kDa protein band for male 6, arguably one of the more abundant bands, peptides identified through cross-species matching had a peak area of less than 5.29 e^5. In contrast, 33 unidentified peptides sequenced *de novo* had a peak area above this, up to 1.89 e^8. The protein band at approximately 16 kDa was not confidently and consistently identified either, although again, inspection of high-quality *de novo* sequences had homology to the secretoglobin family (*C.griseus*).

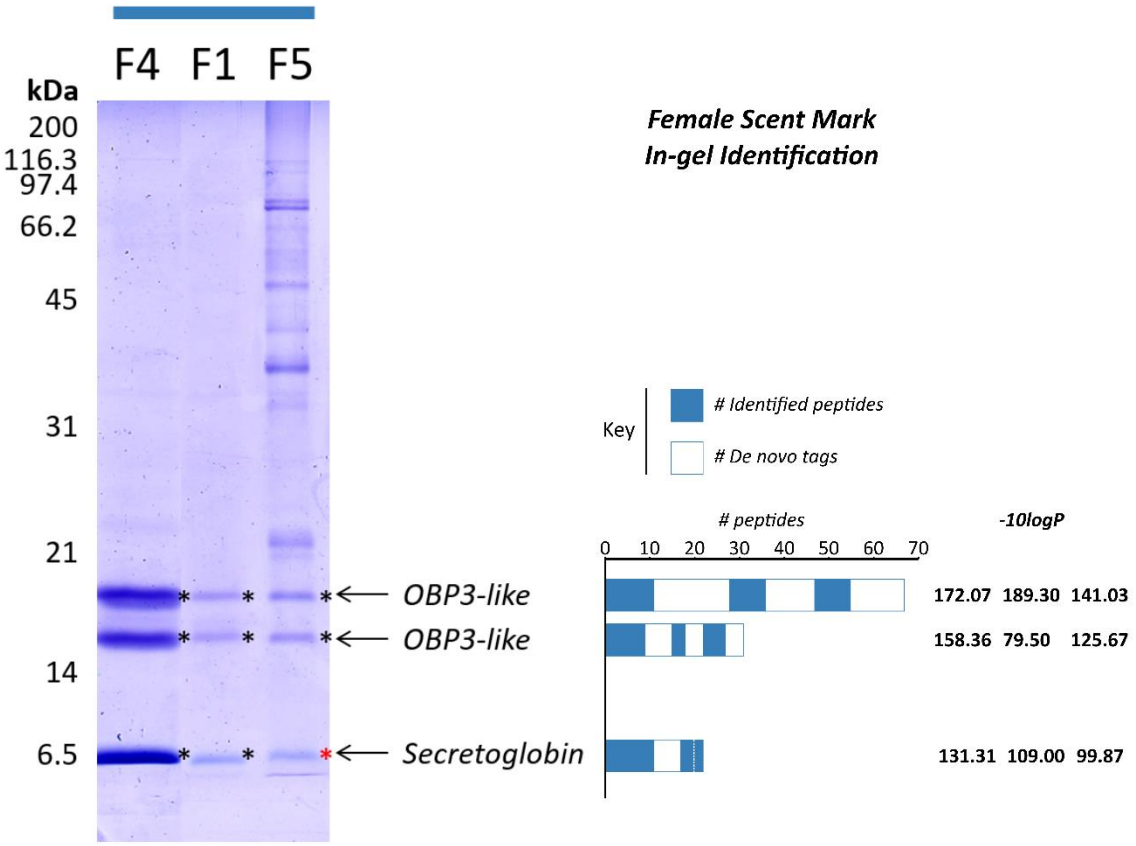


Figure 4.26 | Identification of major female field vole scent mark proteins separated by molecular weight using SDS-PAGE.
Protein in scent marks from female field voles was separated by SDS-PAGE and subject to in-gel digestion with trypsin. LC-MS/MS and subsequent analysis in PEAKS^{TM} suggested likely candidates. The number of identified peptides, through both peptide spectral matching (blue) and *de novo* tags (white), were inspected, in addition to -10logP confidence scores. The number of identified peptides for each excised band are stacked onto the same graph, likewise the -10logP score is given for each replicate.

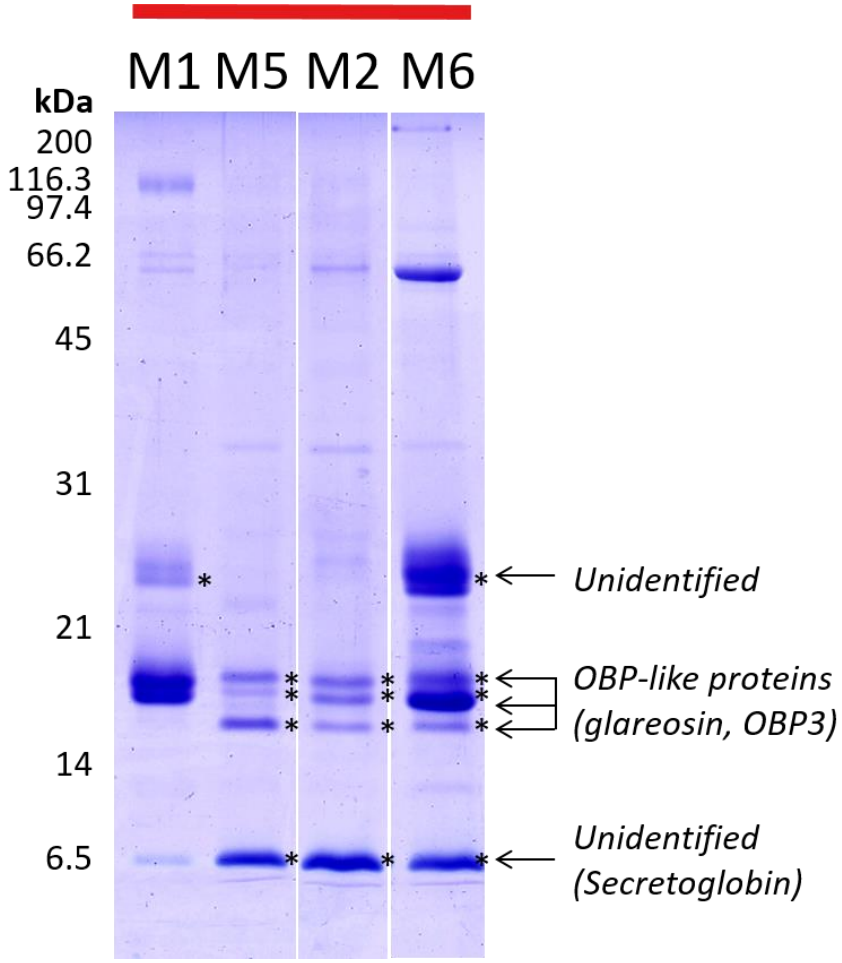


Figure 4.27 | Identification of major male field vole scent mark proteins separated by molecular weight using SDS-PAGE.
Protein in scent marks from male field voles was separated by SDS-PAGE and subject to in-gel digestion with trypsin. LC-MS/MS and subsequent analysis in PEAKS^{TM} suggested likely candidates.

In-gel digestion and subsequent LC-MS/MS analysis revealed that whilst there is strong evidence for OBP-like proteins expressed in the scent marks of both male and female field voles, the picture is far from complete. Identification of proteins in female field vole scent marks was comparatively straightforward. The higher molecular weight proteins were assigned as likely sequence variants homologous to bank vole OBP3. The less confident identification of field vole glareosin peptides is also notable, given that in bank voles, glareosin is male-specific. Prostatic steroid-binding protein is also a major component of female scent mark protein complement, that in bank voles appeared to be male-specific. As discussed in the previous chapter, secretoglobin family 1D member 2, or prostatic steroid-binding protein C1, is a major seminal protein in the rat, and is expressed at its highest level in the testis ((Bastian *et al.*, 2008)). Although the function of prostatic-steroid binding protein has not yet been fully determined, there is evidence of an immunosuppressant role

(Maccioni, Riera and Rivero, 2001). Notably, the equivalent resolved band in males was not identified by cross-species matching, but identification required manual inspection of unidentified spectra. This suggests that the regions of prostatic steroid-binding protein from which the two *C.griseus* proteins are identified do not provide a confident match in the male protein band, and it is possible this is due to different secretoglobin family proteins in males and females.

Genetic differences between species are augmented under evolutionary pressure for mate selection and breeding advertisement and as a result, cross-species matching in matrices with proteins involved in these functions is not always possible.

4.4.8.3 Discovery proteomics of field vole scent marks

To gain more in-depth identifications, the protein complement of scent marks from five male and three female (n = 5) field voles were subject to digestion with trypsin, and analysed by LC-MS/MS. Corresponding SDS-PAGE analysis of these samples can be found in Figure 4.24. The five available female scent marks were analysed, in addition to male 1 (lane 3), male 2 (lane 7), male 3 (lane 13), male 5 (lane 10) and male 6 (lane 12). Data of scent marks were searched against the database of all UniProt sequences in the taxonomic lineage of *Rodentia*, in addition to the sequenced field vole and bank vole glareosin, field vole MUP and OBP3. Data from males and females were processed separately, as label-free quantification was not performed; the aim was for identification only, considering the difficulties with cross-species identification during analysis of in-gel data.

The low confidence in-gel identifications in male field vole scent marks were indicative of missing database information from which to identify the bands. Peptides sequenced *de novo* by PEAKS™ were inspected, and large unidentified peptides that were abundant and confidently sequenced *de novo* were searched against all *Rodentia* sequences in the NCBI database using BLAST. The protein epididymal-specific lipocalin-9-like from *M.auratus* (NCBI Acession XP_012978714) was identified from a peptide 23 amino acids in length with an average local confidence (ALC) score of 90% as HTDESPVPVQQGFSSDQVTGAWK. The protein match had an e-value of 0.005 and an identity score of 70% when compared to the *M.auratus* protein. Three other lipocalin sequences were also identified from searching in BLAST, all named trichosurin from *M.ochrogaster* and two from *C.griseus* (XP_013210978, EGW10042 & XP_027241486, respectively). The epididymal-specific lipocalin-9-like protein was used to also search the UniProt database to find homologous proteins from which to build an alignment of sequences. The top 10 scoring protein sequences were aligned with the epididymal-specific lipocalin 9-like sequence and NCBI sequences in Clustal Omega (Sievers *et al.*, 2011) after removal of the signal peptide in the SignalP server (Petersen *et al.*, 2011) and displayed

in JalView (Waterhouse *et al.*, 2009). Sequences conserved within these protein sequences were then used to search unidentified peptides sequenced *de novo* in male scent mark data. Homologous tryptic peptides could then be identified and aligned (Figure 4.28). The 6 supporting peptides identified were concatenated into a representative protein entry and inserted into the database of *Rodentia* and novel protein sequences to assess quality of identification in comparison to other novel sequences. Due to the discrepancy in terminology regarding the homologous sequences, which will be discussed in the next chapter, the concatenated sequence was referred to as field vole lipocalin 11-like. It is worth noting that the inclusion of this peptide combination simply represented another type of lipocalin as present in the samples, and the extent of heterogeneity in relation to lipocalin 11-like proteins in field vole scent marks is far from understood.

Established lipocalin sequences were identified in both female and male scent mark samples (Figure 4.29 and Figure 4.30, respectively). In females, the OBP3-like draft sequence proposed from analysis of juvenile field vole samples was most strongly identified with a high -10logP score of 219.74. OBP2 (*M.glareolus*), and the lipocalin 11-like peptides were all also strongly identified, although field vole glareosin was not matched. Field vole MUP draft sequence was identified in only one sample. Proteomic analysis of male scent marks most confidently identified the glareosin variants. The other lipocalins (OBP3-like, OBP2 [*M.glareolus*], lipocalin 11-like and MUP) were also identified in male scent marks, although only two samples matched one peptide each of the MUP protein.

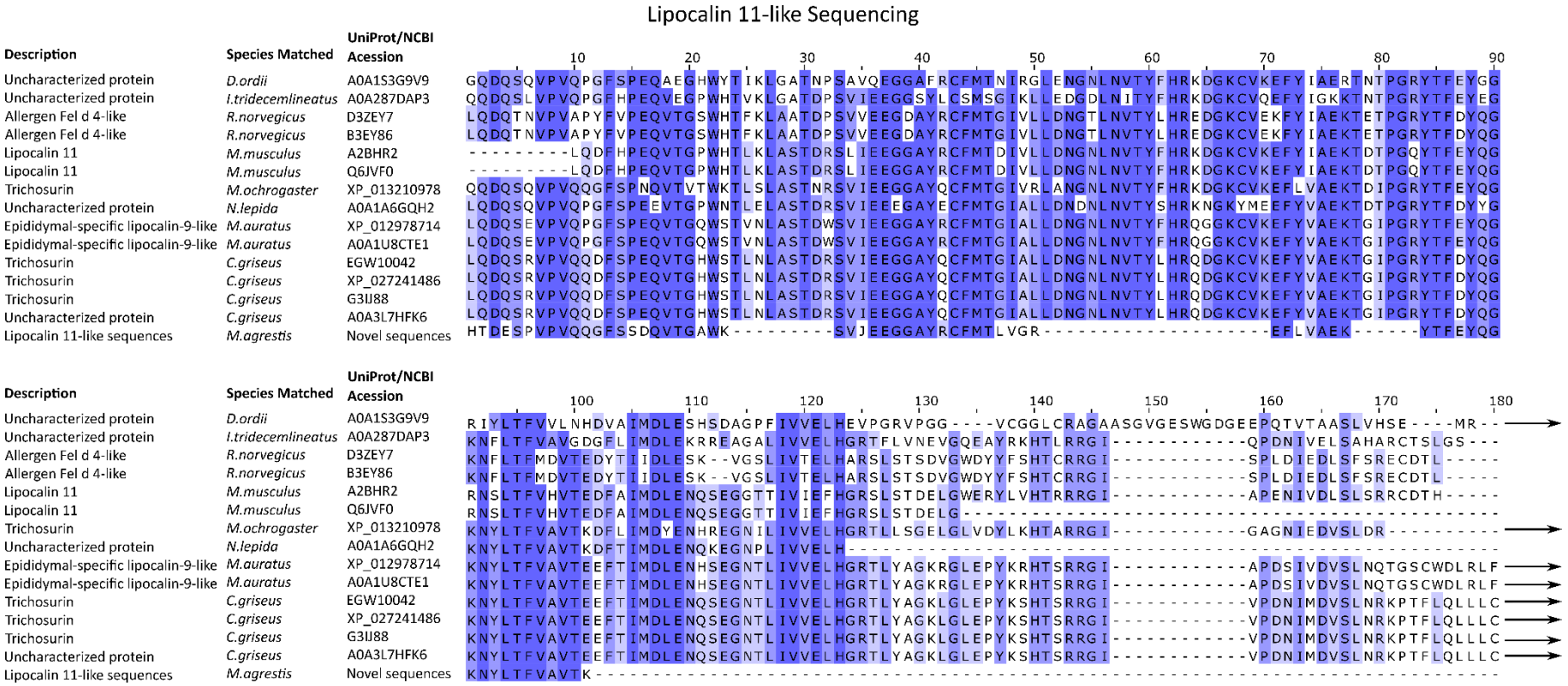


Figure 4.28 | Peptide coverage of lipocalin 11-like proteins observed in male scent marks.
Inspection of confident, abundant peptides sequenced *de novo* from peptide information of male scent marks revealed the presence of sequences homologous to another type of lipocalin. High quality, unidentified peptides were searched using BLAST in the NCBI database of non-redundant protein sequences from *Rodentia*. Epididymal-specific lipocalin-9-like from *M.auratus* was the closest match found. Protein sequences homologous to the *M.auratus* sequence were identified in UniProt, and together with other lipocalin sequences identified from individual peptide searches, signal peptides were removed using the Signal P server (Petersen *et al.*, 2011) and an alignment was generated in Clustal Omega (Sievers *et al.*, 2011) before formatting in JalView 2.0 (Waterhouse *et al.*, 2009).

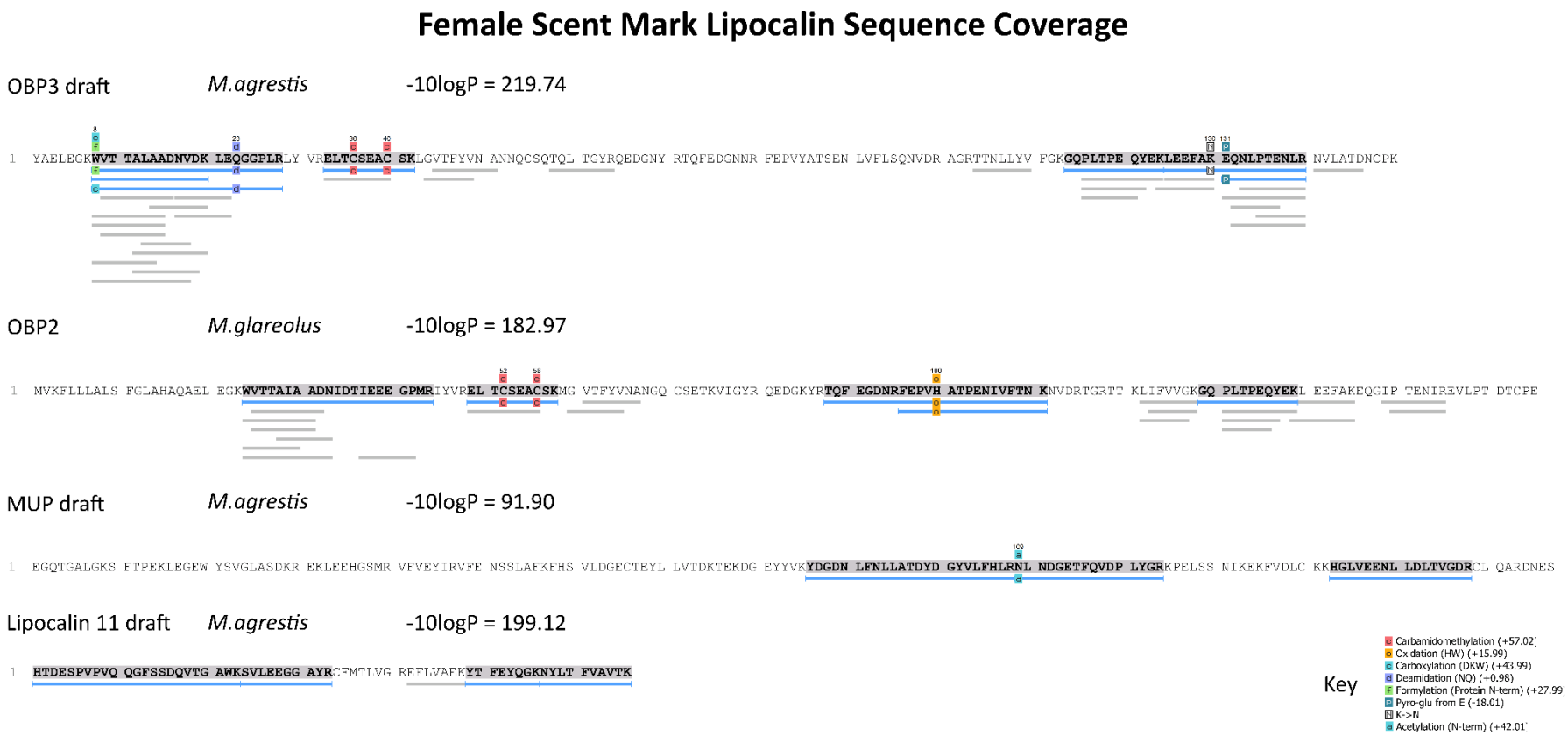
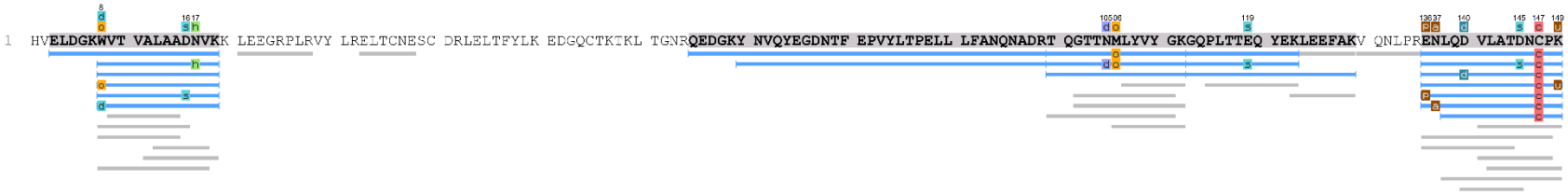


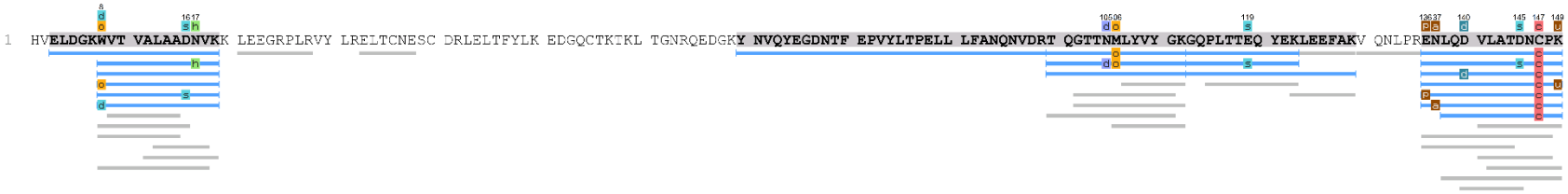
Figure 4.29 | Sequence coverage of lipocalin sequences identified in female scent marks.
Protein from the scent marks (n = 5) from three female field voles were digested by trypsin and analysed by LC-MS/MS. Lipocalin proteins were identified from SPIDER searches in PEAKS^{TM} against *Rodentia* proteins in UniProt and previously established lipocalin sequences.

Male Scent Mark Lipocalin Sequence Coverage (1)

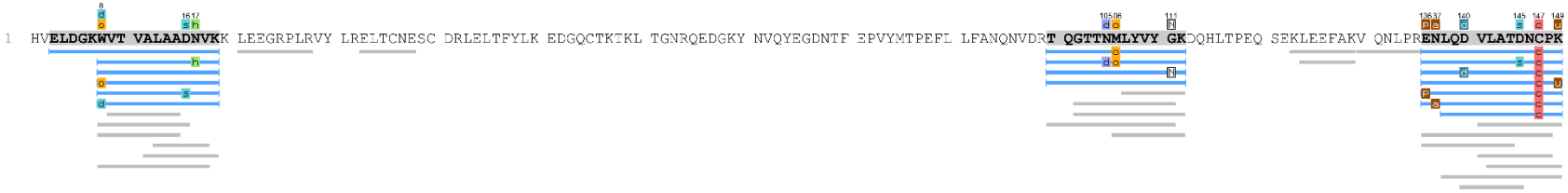
Glareosin (17143 Da) *M.agrestis* -10logP = 414.64



Glareosin (17171 Da) *M.agrestis* -10logP = 383.14



Glareosin (17241 Da) *M.agrestis* -10logP = 358.51



Key

- Ammonia-loss (N) (-17.03)
- Carbamidomethylation (+57.02)
- Deamidation (NQ) (+0.98)
- Dehydration (-18.01)
- Dihydroxy (+31.99)
- Hydroxymethyl (+30.01)
- Oxidation (M) (+15.99)
- Oxidation (HW) (+15.99)
- Pyro-glu from E (-18.01)
- Sodium adduct (+21.98)
- Ubiquitin (+114.04)
- G->N
- K->N
- ISD (z+2)-series (-15.01)

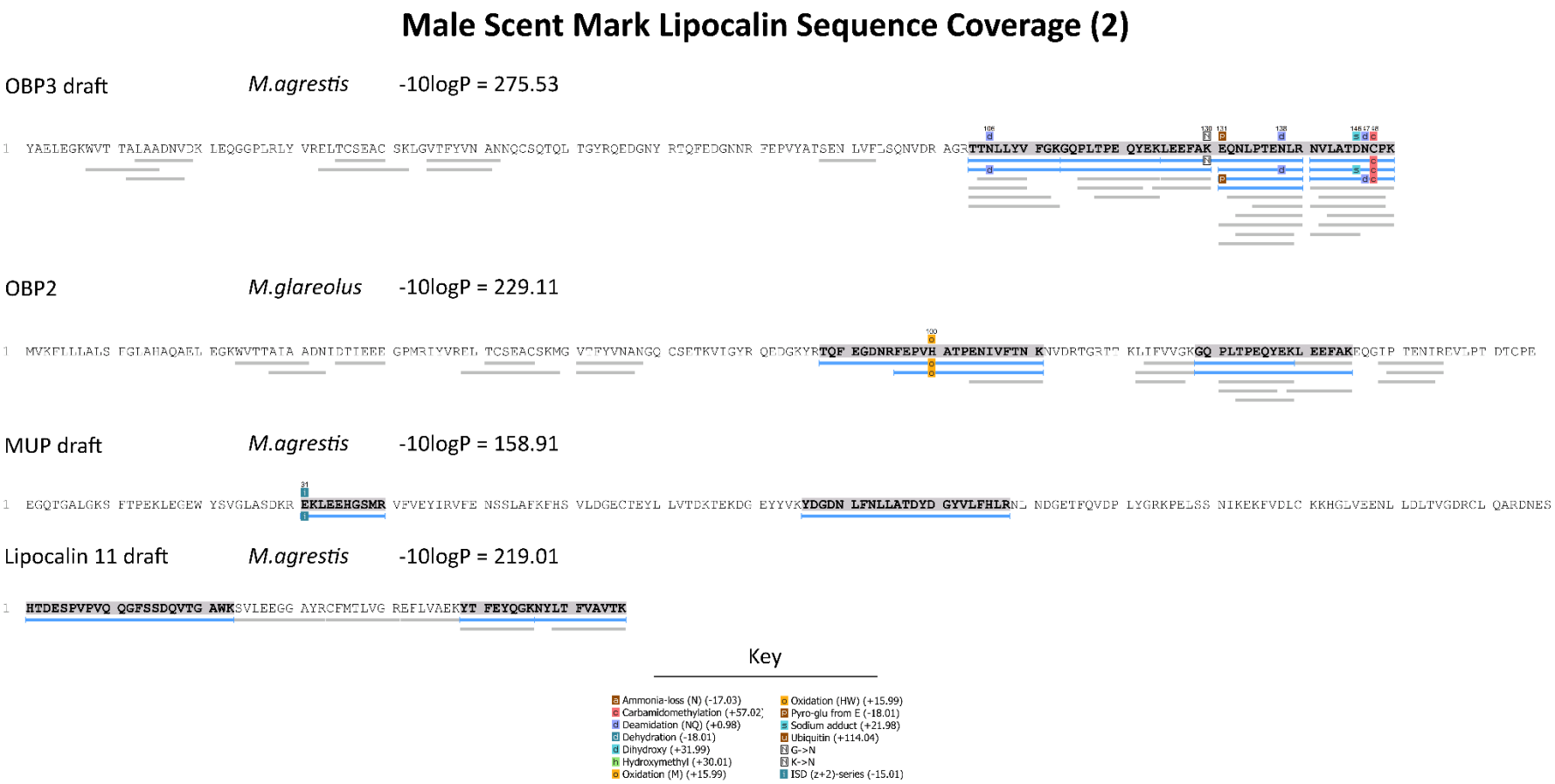


Figure 4.30 | Sequence coverage of lipocalin sequences identified in male scent marks.
Protein from the scent marks (n = 5) from five male field voles were digested by trypsin and analysed by LC-MS/MS. Lipocalin proteins were identified from SPIDER searches in PEAKS™ against *Rodentia* proteins in UniProt and previously established lipocalin sequences.

During inspection of peptides sequenced *de novo*, either unidentified or partially identified, many of the most abundant peptides were noted as homologous to the secretoglobin family, identified by cross-species matching to the *C.griseus* proteins mentioned previously (UniProt Accessions G3IM86 and G3IM12). For example in the top 20 scoring peptides sequenced *de novo* by the software with the highest peak area in male scent marks (Table 4.2), seven are matched to the secretoglobin family proteins, and one matched to an OBP-like protein. Most of the other abundant peptides were highly similar to each other, variations of the peptide SYAPJPYDQK. It was noted that these peptides shared similar features to the peptides homologous to the secretoglobin family: an aromatic amino acid close to the N-terminal, a proline-rich central region and a C-terminal contains glutamic and aspartic acid residues. It was therefore hypothesised that these unidentified peptides also belonged to a member of the secretoglobin family, and individual peptides were searched against non-redundant *Rodentia* proteins in the NCBI database using BLAST. Whilst no identifications of significance were made, the closest matches were members of the secretoglobin family in *M.ochrogaster* (NCBI Acessions XP_005352044 and XP_026637021). An alignment of matched secretoglobin family proteins was constructed, including the *C.griseus* proteins from UniProt, in Clustal Omega (Sievers *et al.*, 2011) after removal of signal peptides using the Signal P server (Petersen *et al.*, 2011) and the resulting alignment was viewed in JalView (Waterhouse *et al.*, 2009). From this alignment, additional secretoglobin sequences were able to be identified. Identified novel sequences also showed a level of heterogeneity, and a number of variants were observed. Three partial sequences comprising the variant sequences observed were constructed for re-searching and these were aligned with secretoglobin sequences (Figure 4.31). No relevance is given to which peptides were concatenated with peptides from other regions of the sequence as no overlapping peptides were available, and constructed sequences were for identification only.

Constructed sequences of secretoglobin-like peptides were inserted into the database and scent mark peptide data were re-searched. Peptides were observed in both female and male scent mark data (Figure 4.32 and Figure 4.33, respectively).

Table 4.2 | Top 20 most abundant peptides sequenced de novo in PEAKS™ from analysis of male scent marks.
Peptides from male scent marks were sequenced in PEAKS™ and manually inspected.

Peptide	*ALC (%)*	*length*	*m/z*	*z*	*Area*	*Accession*
SYAPLPYDQK	95	10	591.296	2	6.93E+07	
YNAPPEAVEAK	98	11	594.7987	2	6.33E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
SYAPFPYDEK	96	10	608.7803	2	3.49E+07	
YSAPFPYN(+.98)EK	89	10	608.7808	2	3.49E+07	
YNAPPEAVEAK	99	11	594.8004	2	3.13E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
EQNLPTENLR	86	10	607.3135	2	3.11E+07	OBP3 (*M.agrestis*) D3VW64 (*M.glareolus*) D3VW63 (*M.glareolus*)
TYAPFPYDEK	92	10	615.7896	2	2.95E+07	
DYAPPEAVEAK	76	11	595.2916	2	2.82E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
SYAPLPYDQK	98	10	591.2969	2	2.19E+07	
TYAPFPYDEK	97	10	615.7877	2	2.11E+07	
YTAPFPYDEK	90	10	615.7892	2	2.11E+07	
SYAPLPYDQK	97	10	591.2973	2	2.10E+07	
NYAPPEAVEAFLK	73	13	724.8763	2	2.10E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
YNAPPEAVEAK	99	11	594.7997	2	2.08E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
YNAPPEAVEAK	99	11	594.7997	2	2.06E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
YDAPPEAVEAK	99	11	595.2911	2	1.79E+07	G3IM86 (*C.griseus*) G3IM12 (*C.griseus*)
DYAPEPADLAK	60	11	595.2907	2	1.79E+07	
TYAPFPYDEK	99	10	615.7885	2	1.65E+07	
GPVC(+57.02)YAVK	99	8	447.2316	2	1.65E+07	
YTAPFPYN(+.98)EK	88	10	615.7892	2	1.65E+07	

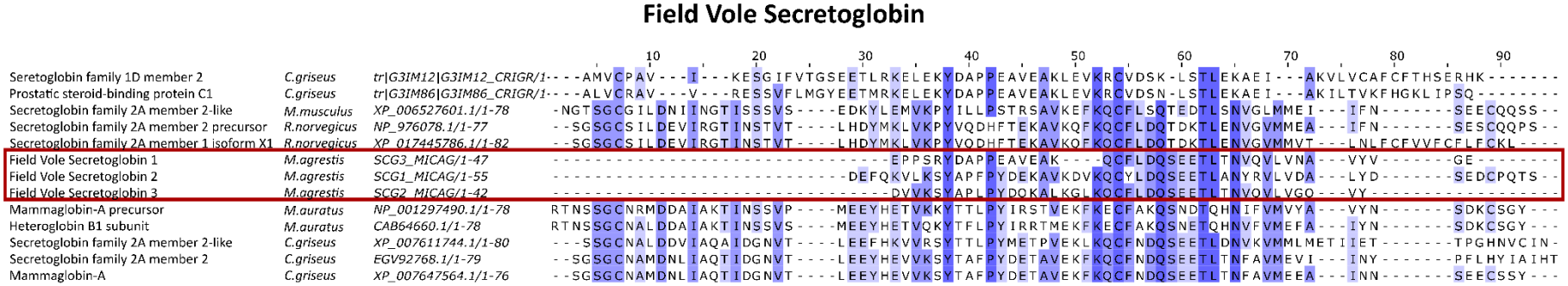


Figure 4.31 | Alignment of rodent secretoglobin sequences with manually selected peptides sequenced in PEAKS™ from analysis of field vole scent marks. The highest abundant peptides in digested protein samples from scent marks obtained from male and female field voles were homologous to the secretoglobin family.

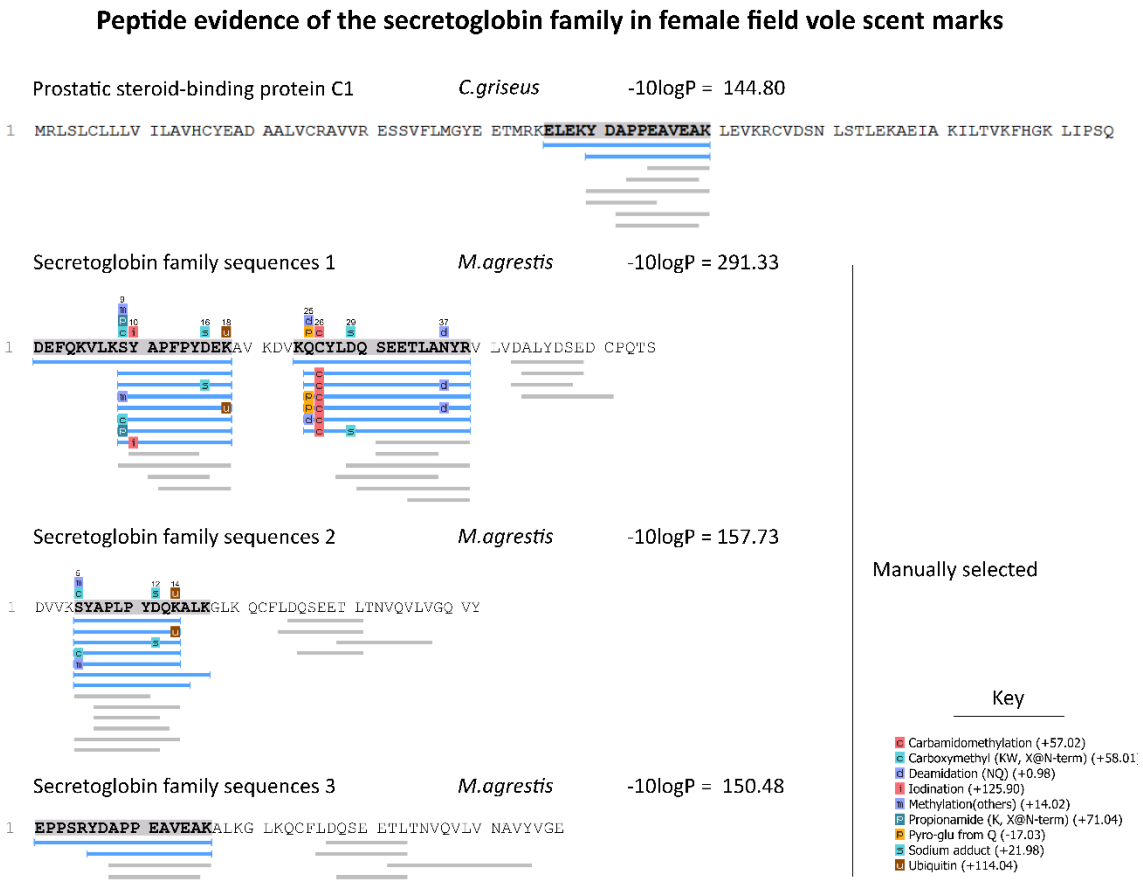


Figure 4.32 | Peptide evidence of the secretoglobin family in female field vole scent marks.
Scent mark peptide data from female field voles were searched against a database that included three sequences constructed from peptides homologous to members of the secretoglobin family.

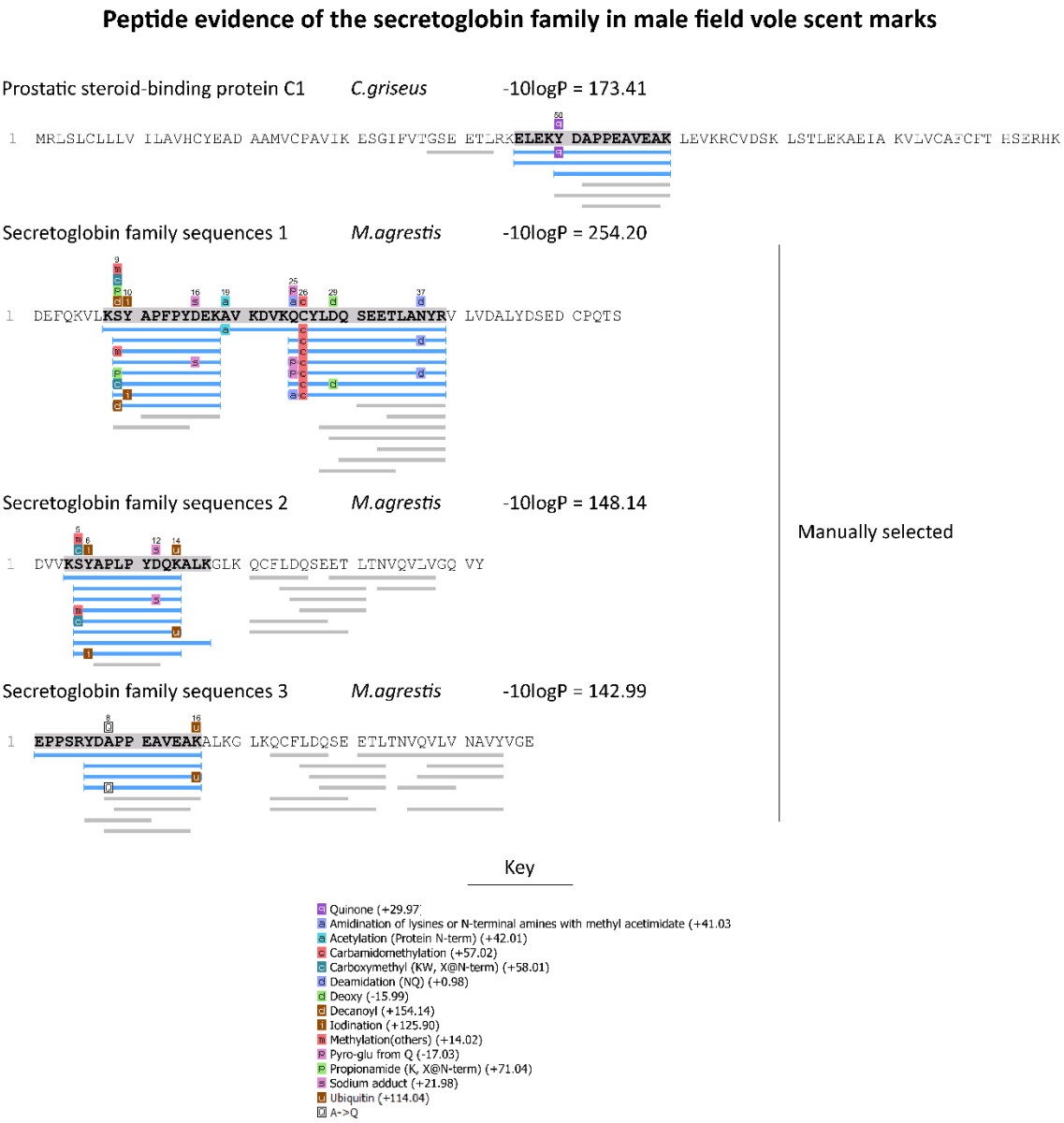


Figure 4.33 | Peptide evidence of the secretoglobin family in male field vole scent marks.
Scent mark peptide data from male field voles were searched against a database that included three sequences constructed from peptides homologous to members of the secretoglobin family.

4.5 Discussion

Investigation into the protein content of field vole urine and scent marks reveals considerably more complex profiles than in bank voles.

The urinary proteome of mature field voles was assessed, revealing a statistically significant sexual dimorphism in terms of overall protein output. Males expressed proteins of approximately 17 kDa, which were later determined to be closely related to bank vole glareosin. However, rather than the single dominant peak observed in bank vole urine, four consistent peaks were observed from protein profiling; 17138, 17168, 17236 and 17252 ± 1 Da. A multiple enzymatic approach was employed to sequence these proteins and identify the mutations at the root of the mass differences. The proposed mutations were mainly located in a central region of the protein, and after structural homology modelling of each, revealed these were all outwardly-facing, occurring mostly at the closed end of the lipocalin. The four putative field vole glareosin variants could be divided into two pairs, members of each pair differing by only one amino acid, and the two pairs differing from each other by six amino acids substitutions. Interestingly, the substituted amino acids in the higher molecular weight pair contributed to a more hydrophobic externally-facing surface. The outward-facing heterogeneity is particularly interesting, as internal variations in mice MUPs change the ligand-binding capabilities of the protein cavity, whilst the heterogeneity in field vole glareosin does not affect the physiochemical properties of the internal calyx. The increase in surface hydrophobicity may reflect a small change in solubility, or indicate an alternative site on the protein that may play an important role in binding and chemical communication. One discrepancy not solved was the origin of the mass peak 17252 Da. Two explanations were possible from the dataset; an amino acid mutation JTC[N>K]ESCDRJE, or oxidation of a methionine residue, for both of which there was peptide evidence, however this could not be confidently elucidated. Evidence of other, lower abundant OBP-like peptides were also present in field vole urine, and the mutated peptide may come from another related protein. More stringent purification may be able to discriminate between these possibilities, but whilst the underlying complexity of these samples is not fully determined, this will be difficult to elucidate.

In addition to field vole glareosin, peptide evidence of a major urinary protein-like sequence was discovered. Whilst its presence was later established to be highly abundant in both male and female field vole urine (although determining relative abundance of MUP and field vole glareosin was difficult due to the lack of unique peptides), it was discovered in female urine, and is proposed to explain the dominant mass peaks in females of approximately 21 kDa. The protein(s) was partially sequenced *de novo*, with the exception of a central portion, which was hypothesised to be

concealed by a glycan residue. A discrepancy between the sequence mass and observed mass of 2 kDa and the absence of sequence coverage available in a region commonly glycosylated in homologous proteins (Mechref *et al.*, 2000; Scaloni *et al.*, 2001; Mastrogiacomo *et al.*, 2014) suggests that this is a strong possibility, and a future step would be to complete the sequence in a similar way to the approach taken in chapter 5, in addition to determining the glycan moiety. Another aspect not considered is the heterogeneity of MUPs present. This is the first indication of a major urinary protein in the urine of a *cricetidae* member. It has been thought that MUPs were a feature of murid rodents, whilst secretory lipocalins in the cricetid family tended to be OBPs (Briand, Trotier and Pernollet, 2004; Stopková *et al.*, 2010b; Turton *et al.*, 2010; Loxley *et al.*, 2017). The lipocalin profile of the field vole therefore seems to provide evidence of another signalling mechanism.

Inspection of peptides at lower abundance suggests further complexity. In-gel identification provided evidence of peptides homologous to the bank vole protein OBP2 in females, and additional OBP-like peptides were commonly seen during manual inspection of unidentified peptides.

Analysis of mature field vole urine also contained good peptide evidence for a novel field vole OBP3 and a small level of evidence for an epididymal-specific lipocalin from cross-species matches to homologous protein of *C.grieus* and *M.musculus* (UniProt Accessions G3I936_CRIGR and NPC2_MOUSE, respectively), indicative of another type of lipocalin present. Global proteomics also identified both prostaglandin-H2 D-isomerase, which in mice is thought to be involved with the maturation of sperm (Gerena *et al.*, 2000) and prostatic steroid-binding protein (and potentially other members of the secretoglobin family). In the rat, this protein is expressed most abundantly in the testis (Bastian *et al.*, 2008). However, both of these proteins, which were identified exclusively in males when analysing bank vole urine, were confidently identified in both male and female field voles urine.

Preliminary analysis of proteins in a small number of juvenile animal urine samples, whilst relatively inconsistent compared with the strong 17 kDa profiles of mature males, provided evidence that juvenile field voles also express urinary proteins, and may utilise this established high molecular weight mechanism for communication with conspecifics. Despite variability in intact mass profiles, discovery proteomics identified consistent expression of the novel field vole MUP sequence, and some level of expression of field vole glareosin in some males. OBP3-like peptides were also assembled to form a draft sequence for field vole OBP3, potentially the origin of the consistently observed 16893 Da mass.

Scent marks were also highly proteinaceous. Not enough samples were available to determine if sexual dimorphism occurred at the overall protein level, however SDS-PAGE analysis provided evidence of a band at approximately 17 kDa exclusively observed in males, which is likely to be glareosin. The remaining protein bands at the lower molecular weight region were consistent between males and females, however. Peptide evidence for members of the secretoglobin family were seen in both males and females. Likewise other protein bands were identified from peptides homologous to OBPs. Identification of protein bands was especially difficult in male scent marks, and many peptides sequenced were left unidentified. Manual inspection identified another type of lipocalin, named here as lipocalin 11, although terminology of the most closely related proteins also included trichosurin, epididymal-specific lipocalin 9-like and allergen fel d 4-like. Terminology for unannotated proteins such as these is often unclear and incorrect. Phylogenetic analysis of these proteins would reveal more about their origins as lipocalin sequences, however this approach was not pursued due to the fact only a small number of peptides could be identified. Discovery proteomics of scent marks also revealed highly heterogeneous peptides belonging to the secretoglobin family in both males and females, and appears to be more of a feature in the scent mark proteome than in the urinary proteome.

The presence of novel sequences and notably identified proteins are summarised in Table 4.3. Proteomic-based analysis of field vole urine and scent marks reveals a far more complex landscape than in the bank vole, and in terms of types of lipocalins identified, possibly more complex than the MUP heterogeneity of mice and rats. Strong peptide evidence has been submitted for not only variants of field vole glareosin, but also an abundant OBP3 sequence, a field vole major urinary protein and another type of lipocalin, possibly related to epididymal-specific lipocalin 9 (salivary lipocalins). A great deal more work is needed to fully elucidate not only the level of complexity and heterogeneity of these proteins present in these secretions, but also their potential role as odour cues. Just as their behavioural biology is complex, the possible underlying mechanism of the associated social behaviour is too, and separating out the roles for each of these lipocalin types could provide an in-depth perspective on their behavioural mechanisms. However, a genome-free proteomics approach for this would require teasing apart many different proteins at low abundances with high levels of similarity. Whilst a genome-free approach pulls into sharp focus missing portions of sequence that could indicate a post-translational modification that may be missed without the detailed and manual consideration taken here, understanding lower level complexity is unlikely to be completely achieved without a comprehensive database.

Table 4.3 | Summary of novel and established lipocalin sequences and other notable protein identified in field vole urine and scent marks.
Novel proteins were re-searched in previous datasets to assess presence in urine and scent marks, of male and female and from mature and juvenile animals.

	Protein	Sequence species	Urine				Scent Marks		Sequencing stage	Comments
			Mature		*Juveniles*					
			Male	Female	Male	Female	Male	Female		
Novel	Glareosin (3 variants)	*Microtus agrestis*	✓	X	✓	X	✓	X	Fully sequenced	Exception is variant 17252 Da- not enough evidence to support amino acid substitution over oxidation)
	MUP	*Microtus agrestis*	✓	✓	✓	✓	✓ *	✓ *	Partial sequence	Suspected glycosylation site. Heterogeneity not assessed.
	OBP3-like	*Microtus agrestis*	✓	✓	✓	✓	✓	✓	Draft sequence	Sequence not confirmed with equivalent intact mass peak. Heterogeneity not assessed.
	Lipocalin-11	*Microtus agrestis*	✓ *	✓ *	✓ *	✓	✓	✓	6 peptides	Full sequence not possible; heterogeneity of this type of lipocalin unknown
	Secretoglobin	*Microtus agrestis*	✓	✓	✓	✓	✓	✓	Multiple homologous peptides.	Strong evidence for heterogeneity
Cross-species matching	OBP3	*Myodes glareolus*	✓	✓	✓	✓	✓	✓	Cross-species match	
	OBP2	*Myodes glareolus*	✓	✓	✓	✓	✓	✓	Cross-species match	
	OBP1	*Myodes glareolus*	X	X	✓ *	✓ *	X	X	Cross-species match	
	Epididymal secretory lipocalin	Multiple	✓	✓	✓	✓	X	X	Cross-species match	
	Prostaglandin H2 D-isomerase	Multiple	✓	✓	✓	✓ *	X	X	Cross-species match	

*identified in less than half of the samples searched

4.6 Supplementary

Below is a comprehensive list of relevant information available in the supplementary material.

4.6.1 Intact mass profiles of urine from mature field voles.

4.6.2 Fragment ion spectra of peptides used to sequence an initial draft of field vole glareosin.

4.6.2.1 Initial sequence

4.6.2.2 Mutations

4.6.2.3 Peptide map of glareosin variants for individual males with distinctive protein profiles, used to confirm mutation sequencing.

4.6.3 Sequence coverage of odorant binding proteins from analysis of pooled female field vole urine.

4.6.4 Alignment of major urinary proteins to support preliminary sequencing of a field vole major urinary protein.

4.6.5 Intact mass profiles of juvenile field vole urine.

5 Characterisation of the urinary protein content in the New Zealand brushtail possum, *Trichosurus vulpecula*

Abstract

The brushtail possum, *Trichosurus vulpecula*, is a marsupial species native to Australia, but since its introduction to New Zealand in the 1800s has thrived as a pest species, causing catastrophic damage to the native flora and fauna. Further research into the biology of this species is required to manage populations effectively. Proteins have been established to play an important role in chemical communication, and developments in pest management have targeted pest species' olfactory communication as a means of control. The common brushtail possum (*Trichosurus vulpecula*) exhibits scent marking behaviours that use urine, and consequently the protein component of male and female brushtail possum urine was investigated. It contains a prominent glycosylated protein of approximately 20 kDa predominantly expressed in males. This protein has been isolated and its sequence completely determined by mass spectrometry including the use of metabolic labelling to resolve leucine/isoleucine isobaric ambiguity. The protein was identified as a lipocalin, and phylogenetic analysis grouped the protein with other marsupial lipocalin sequences in a phylogenetic clade distinct from other cross-species lipocalin types. The pattern of expression in male possum urine and the similarity in sequence and structure to other lipocalins suggests this protein may have a role in brushtail possum chemosignalling.

Contributions

This project was in collaboration with the research group of Professor Wayne Linklater (School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand). Urine sampling was performed by Dr David Hooks (Victoria University of Wellington, New Zealand). Anion exchange chromatography and prior sample preparation of pooled male urine, including SDS-PAGE and protein and creatinine assays, were jointly carried out by Grace Loxley and Dr David Hooks at the University of Liverpool. Statistical analysis of protein and creatinine assay data was undertaken by Professor Jane Hurst (Mammalian Behaviour & Evolution Group, Institute of Integrative Biology, University of Liverpool, Liverpool, UK).

5.1 Introduction

The common brushtail possum, *Trichosurus vulpecula* (Kerr, 1792) is a marsupial native to Australia. A protected wildlife species in mainland states (Wildlife Act, 1975), this species is of Least Concern (LC) on the International Union for Conservation of Nature Red List (IUCN, 2018). However, in some urban and agricultural regions, possums are present in high densities and population management is required; an annual cull is permitted to protect agricultural crops in Tasmania, for example (Wildlife Regulations 1999). Since its introduction to New Zealand in the mid-19th century to develop a fur industry (Pracy, 1962), possums have become a major pest. The species has been the subject of a huge conservation effort by the New Zealand government, which aim to eradicate the brushtail possum by 2050 (Cabinet, 2016). A highly adaptive aboral browser, the brushtail possum destroys native flora and fauna (Fitzgerald, 1976; Allen, Fitzgerald and Efford, 1997; Fitzgerald and Gibb, 2001; O'Donnell, Weston and Monks, 2017), and predates on native bird and invertebrate species (Brown, Innes and Shorten, 1993; Innes *et al.*, 2005), which since the isolation of the landmass 80 million years ago, have evolved without the threat from almost all mammals and are therefore ill adapted to resist predation. It is also a reservoir and vector for bovine tuberculosis (*Mycobacterium bovis*), taking a considerable economic toll on the cattle and deer industry (Morris and Pfeiffer, 1995; Caley *et al.*, 1999; Corner, 2006).

Current pest control methods are largely dominated by use of sodium monofluoroacetate, commonly known as "1080" (Eason *et al.*, 2017). Widespread distribution by air, used on about a fifth of the control area, has attracted opposition (Green and Rohan, 2011), although there is little evidence of long-term adverse environmental effects (O'Halloran *et al.*, 2005; Srinivasan and Suren, 2018), and a positive long-term response of native bird species populations (Greene *et al.*, 2013; Kemp *et al.*, 2018; Van Vianen *et al.*, 2018). Research into biological control has come into focus (Barlow 1994; Ramsey 2005), but faces a knowledge gap when it comes to behaviour and physiology of marsupial species in general, and the brushtail possum specifically.

The brushtail possum is largely herbivorous and solitary, with a lifespan of approximately 13 years (Thomson, 1921). It matures at one year and, under favourable conditions, females can raise a single pouch young twice a year (Efford 1998); the primary breeding season takes place in autumn but breeding behaviour appears to be flexible and dependent on environment (Jolly, Scobie and Coleman, 1995). Possums are generally solitary animals with small, overlapping home ranges (Crawley, 1973; Ward, 1984), rarely interacting except

during the breeding season. Interactions usually only occur between two individuals in the overlapping regions (Ji, White and Clout, 2005). Males tend to have the larger home ranges, within which they move more frequently, than females, and for males dominance is correlated with age and size (Brown & Macdonald 1985). Possums exhibit social behaviours observed in many other mammals, involving social hierarchy, territorial defence, individual recognition and mating (Russell, 1985).

Dominant males, and occasionally dominant females, exhibit a number of scent marking behaviours including chinning, chesting, cloacal dragging, urine dribbling and face-washing (McLean 2014; Spurr & Jolly 1999). The purpose of these behaviours has yet to be established, as does the physiological and molecular mechanism behind them. Although no pheromones have been identified to date, possums display scent marking behaviour which involves rubbing their chin, sternum, or cloaca on surfaces (Spurr and Jolly, 1999). They also have an extensive olfactory epithelium and a well-developed vomeronasal organ, both of which indicate an advanced olfactory capability (Brown & Macdonald 1985).

The sternum patch is particularly distinctive in sexually mature males and is due to the presence of active apocrine and sebaceous glands, secretions of which are composed of fatty acids, small heterocyclics, phenolics, and aldehydes with the major components being waxy lipids (McLean, Davies and Wiggins, 2012). The cloacal glands contain triacyl glycerides (C_{10}-C_{18}) and branched fatty acids (C_7-C_9) (Woolhouse, Weston and Hamilton, 1994; Zabaras, Richardson and Wyllie, 2005). The paracloacal glands, arranged in pairs either side of the cloaca, produce either an oily secretion or cell-containing medium. The oily secretion is a complex mixture of triacyl glycerols, diacylglycerol ethers and triacylglycerol estolides, which have not been reported in animals previously (McLean *et al.*, 2015). In addition to glandular secretions, cell-containing material is continuously secreted into urine and faeces making urine a potential source of possum chemosignalling. Whilst the presence of these glands, and to an extent the secretory components, is established, no correlations have been made with behavioural traits yet. There is also little research so far into the protein components of any of the above scent secretions, but from what we know about the role of proteins in scent communication in other mammals, discussed previously, it is likely that the protein component is worth investigation.

Little is known about chemosignalling proteins in marsupials although available data suggests well-developed vomeronasal organs similar to that in rodents (Schneider et al. 2008; Poran 1998; Aland et al. 2016; Brown & Macdonald 1985). Lipocalin-like proteins are

secreted in milk during lactation of the tammar wallaby (Nicholas *et al.*, 1987; Collet, Joseph and Nicholas, 1989; Trott *et al.*, 2002), red kangaroo (McKenzie, Muller and Treacy, 1983), grey kangaroo (McKenzie, Muller and Treacy, 1983), quokka (Beg and Shaw, 1994) and the brushtail possum (Piotte *et al.*, 1998). Of these lipocalin-like proteins excreted in the brushtail possum, one was homologous to β-lactoglobulins found in the other marsupial species, but the other more closely resembled a major urinary protein (MUP). Trichosurin was the first solved 3D structure of a lipocalin from a metatherian and, compared to other known lipocalins discussed in Chapter 1, had an unusual dimer arrangement, however the function of trichosurin is unknown (Watson *et al.*, 2007). Watson *et al*. (2007) discuss the possibility that the demonstrable ability of the protein to bind small phenolic compounds could be involved in priming the neonate liver to produce the enzymes required to metabolize otherwise toxic plant phenols, thus reducing dietary limitations.

5.2 Aims & Objectives

Due to the importance of the brushtail possum in the future of New Zealand's ecology, a greater understanding of the biology and social organisation of this species is required to target and monitor populations effectively. Brushtail possums have been observed to use urine in scent communication, but the molecular components contributing to the potential signal have not been explored. Urinary proteins are used extensively in other mammalian systems to facilitate olfactory communication. Consequently this project focussed on an assessment of the urinary proteome of the brushtail possum, *Trichosurus vulpecula*, which has not been previously investigated. The following objectives were set for this project;

a. To assess the protein content of brushtail possum urine
b. To isolate and sequence abundant proteins with homology to proteins with a known chemosignalling function.
c. To explore sequence and structural homology with proteins suggested to have a chemosignalling function in other species.
d. To investigate differences in protein expression based on sex.

5.3 Methods

All methods are essentially the same as those detailed in Chapter 2: Experimental Strategy. Any adaptations or additional methods are as follows:

5.3.1 Urine sampling

Urine samples were collected from wild-caught *T.vulpecula* captured in the Wellington region of New Zealand. Possums were kept in separate wooden hutches, provided with dried food ad libitum, and given fresh vegetables and browse weekly. Urine was collected in 50 mL plastic tubes. Samples (n=12) from a first set of 7 males and 5 females were stored and shipped at -20 °C. Further samples (n=30) collected from 9 males and 5 females were freeze dried before shipment and reconstituted in MilliQ-grade water to a final volume ten times less than that of the original. Concentrated urine samples were used to create male and female pooled urine.

5.3.2 Protein and creatinine assays

Full protocols can be found in Chapter 2: Experimental Strategy, Section 2.3.1.

5.3.3 SDS-PAGE and Native PAGE analysis

Full protocols can be found in Chapter 2: Experimental Strategy, Section 2.3.2.

5.3.4 Anion exchange chromatography

General protocols can be found in Chapter 2: Experimental Strategy, Section 2.8.

Samples of pooled male and pooled female urine were dialysed into 20 mM Tris-HCl pH 8.5 using 3.5 MWCO SnakeSkin™ dialysis tubing (Thermo Fisher Scientific) overnight at 4°C. Buffer was then replaced and the sample was dialysed for a further four hours. Centrifugation at 2000 x *g* and the supernatant retained. Protein (350 µg) was then loaded onto a 1 mL RESOURCE Q column (GE Life Sciences) and eluted at 1 mL/min over a linear gradient of 0-0.8 M NaCl in 20 mM Tris-HCl buffer. Manually collected fractions for each detected peak were analysed by SDS-PAGE. The 25 kDa protein eluted from the column in approximately 0.25 M NaCl and fractions containing the protein were pooled for further analysis.

5.3.5 Intact protein mass measurement

Instrument methodology remained unchanged and can be found in Chapter 2: Experimental Strategy, Section 2.5.1. Urine samples were desalted using 7K MWCO 0.5 mL Zeba™ Spin Desalting Columns (Thermo Fischer Scientific) where specified (protocol in Chapter 2: Experimental Strategy, Section 2.9.3) when initial analysis yielded poor quality spectra. Between 5-80 pmol protein was analysed in each sample injection. For intact mass measurement of deglycosylated protein, the following protocol was developed (method development; Section 5.4.5). The protein (6 µg) was diluted in 8 M urea to a final concentration of 5 M urea and incubated at 37 °C for one hour. Dithiothreitol was added to a final concentration of 150 mM and incubated at 37 °C for a further 30 minutes. The protein was deglycosylated using 500 units PNGase F, incubating for one hour at 37 °C. The buffer was exchanged for HPLC-H_2O using Zeba™ Spin Columns and diluted into 95% HPCL-grade H_2O, 5% MeCN, 0.1% formic acid before analysis.

For spectral comparison, mass spectra were aligned in SpecAlign (Wong, Cagney and Cartwright, 2005). Peaks within the region of 20 -21 kDa were aligned using the PAFFT method (Misawa and Ootsuki, 2015) against an averaged spectrum. Intensities were normalised to TIC.

5.3.6 In-gel digestion and MALDI-ToF analysis

Full protocols can be found in Chapter 2: Experimental Strategy, Sections 2.5.1and 2.6.2.

5.3.7 Protein desalting and concentration

For digestion with multiple proteases, protein in the relevant pooled anion exchange fractions was captured using Strataclean resin™, as detailed in Chapter 2: Experimental Strategy, Section 2.9.1. Prior to deglycosylation, the pooled protein-containing fractions were concentrated 10-fold using Vivaspin® columns (Sartorius), the full protocol of which can be found in Chapter 2: Experimental Strategy, Section 2.9.2. For future glycan and intact mass analysis, the buffer was exchanged by diluting 10-fold in 10 mM NaCl, concentrating to the original volume, and repeating twice using Vivaspin® columns (Sartorius).

5.3.8 *In-solution digestion and LC-MS/MS*

No changes were made from that specified in Chapter 2: Experimental Strategy, Section 2.4.2, with the exception of a modified protocol for digestion of deglycosylated protein. The method development of this modified protocol can be found in Section 5.4.5. The final protocol was as follows: The protein (2 µg per digest) was diluted in HPLC-grade H_2O, and Rapigest™ SF Surfactant was added to a final concentration of 240 µM and incubated at 80 °C for 10 minutes. Dithiothreitol was added to a final caoncentration of 0.5 mM and incubated at 60 °C for 10 minutes. The reaction mixture was cooled to room temperature and 500 units PNGase F (glycerol-free, New England Biosciences) per 1 µg protein were added. Following incubation at 37 °C for one hour, the reaction volume was divided into the number of digestions planned and diluted to give a final digestion volume of 25 µL into buffers appropriate for the digestion enzyme in question. The protocol then followed standard in-solution digestion protocols from the point of acetylation with iodoacetamide onwards. Instrument methodology is detailed in Chapter 2: Experimental Strategy, Section 2.5.3).

5.3.9 *Peptide data searches*

Data analysis methods and parameters are outlined in Chapter 2: Experimental Strategy, Section 2.6.1. Data from analysis of brushtail possum urine was analysed in PEAKS™ (Bioinformatics Solutions Inc.). Specific parameters and database information related to the analysis of LC-MS/MS data following proteolytic digestion of urinary proteins from *T.vulpecula* were as follows.

Changes to variable modifications in PEAKS DB searches included:

- Deamidation of asparagine to search for peptides modified by the removal of an N-linked glycan by the enzyme PNGase F.
- Trideuterium-labelled leucine to assist with distinction of leucine and isoleucine residues following the dietary incorporation of L-leucine-5,5,5-d_3.

Searches were performed against databases created from all mammalian sequences available in SwissProt, and all grey short-tailed opossum (*Monodelphis domestica*) protein sequences in UniProt. For each search, the relevant proteases were incorporated into the databases to allow for autolysis products, and for deglycosylated peptide data, the protein

sequence of peptide:N-glycosidase F was included. Once confirmed, the newly sequenced protein was subsequently added.

5.3.10 Label-free quantification

Parameters can be found in Chapter 2: Experimental Strategy, Section 2.6.3. Changes to analysis of brushtail possum samples were as follows. Peptide data were generated from the digestion of all urinary proteins in 7 male and 5 female samples followed by LC-MS/MS. Identifications from PEAKS DB, PEAKS PTM and SPIDER searches in the PEAKS^{TM} platform against a database of *Monodelphis domestica* protein sequences, into which the sequences of proteases used were added, were used at a cut-off score 1% false discovery rate (FDR).

Progenesis QI for Proteomics by Nonlinear Dynamics (Waters, Manchester, UK) was used to explore variation at the feature level, and parameters can be found in Chapter 2: Experimental Strategy, Section 2.6.3. No identification was attempted with *T.vulpecula* urine data.

5.3.11 Protein deglycosylation

Protein in urine or in anion exchange fractions was deglycosylated according to the manufacturer's (New England Biolabs) protocol. Purified protein (0.5 µg) was made up to 10 µL and incubated with 2 µL glycoprotein denaturing buffer (0.5% SDS, 40 mM DTT) for 10 minutes at 100 °C. The protein sample was cooled on ice before mixing with 1 µL 10% NP-40, 1 µL 10X GlycoBuffer 2 (50 mM sodium phosphate pH 7.5) and 1 µL PNGase F solution (glycerol-free) (New England Biosciences). After incubation at 37 °C for 1 hour, 10 µL of the digest was analysed by SDS-PAGE.

Deglycosylation protocols were developed separately for subsequent protease digestion and LC-MS/MS analysis, and for measurement of the deglycosylated protein's intact mass (see section 5.4.5).

5.3.12 Discrimination of leucine and isoleucine by metabolic labelling:

General protocol can be found in Chapter 2: Experimental Strategy, Section 2.7. Isobaric amino acids leucine and isoleucine were discriminated by the addition of L-leucine-5,5,5-d_3 to the diet of one possum over a seven day feeding period. Urine was collected once daily and the remaining food removed and weighed. The added L-leucine-5,5,5-d_3 comprised approximately 40% of the dietary leucine this possum received over the feeding period.

Urine samples were desalted using Zeba^{TN} Desalting spin columns (Thermo Fisher Scientific) before in-solution digestion, as described previously. Samples from each day of the feeding period were digested with trypsin only to monitor heavy leucine incorporation. The desalted urine sample from the third day, determined as the sample with highest levels of heavy leucine incorporation was deglycosylated prior to proteolytic digestion with either trypsin only, glu-C only, or with both asp-N and trypsin, whereby trypsin was added four hours after asp-N, followed by overnight incubation as above. Peptides were analysed by tandem mass spectrometry as described in section 2.6.3.

5.3.13 Phylogenetic analysis

Details for phylogenetic analysis methods can be found in Chapter 2: Experimental Strategy, Section 2.10. Specific methods used for analysis of brushtail possum urine are described forthwith. The finished protein sequence was searched using BLAST® against all marsupial non-redundant protein sequences (taxID:9263) in NCBI database or all marsupial sequences in UniProt. Significant (e-value<0.01) results were included in the analysis. Mammalian proteins belonging to the lipocalin family, previously identified as having a putative chemosignalling function, were also included. A full list of UniProt and GenBank accession numbers for the protein sequences used can be found in Supplementary. Predicted signal peptides were removed using the SignalP 4.1 server (Petersen *et al.*, 2011) and aligned using Clustal Omega (Sievers *et al.*, 2011). The resulting alignment was viewed in JalView (Waterhouse *et al.*, 2009) and MEGA 6.06 (Tamura *et al.*, 2013) was used for evolutionary analyses. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model. Bootstrapping analysis using 500 replicates was carried out and the tree with the highest log likelihood (-16237.8530) is shown. Branches corresponding to partitions reproduced in less than 50% boostrap replicates were collapsed. All positions containing site coverage of less than 95% were eliminated and the resulting 131 positions were analysed.

5.3.14 Homology modelling:

General protocol can be found in Chapter 2: Experimental Strategy, Section 2.10.4. All RCSB Protein Data Bank (PDB) structures were searched against the newly sequenced protein using BLAST® (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The 6 top-scoring sequences were aligned with vulpeculin using Clustal Omega (Sievers *et al.*, 2011) and the corresponding

structures used as templates. The alignments were manually adjusted to fit the structural information given in the .pdb file, and 10 models were generated based on each template using Modeller 9.16 (Šali and Blundell, 1993). Model quality was assessed using MolProbity (Chen *et al.*, 2010) and QMEAN score (Benkert, Silvio C. E. Tosatto and Schomburg, 2008). The highest quality model was viewed and annotated in PyMOL (PyMOL, Version 2.1.0).

5.4 Results

5.4.1 Preliminary investigation into urinary protein content

Urine samples were collected from 10 male (n = 26 samples) and five female (n = 16 samples) common brushtail possums. To provide an overall assessment of protein complexity, SDS-PAGE analysis was permorfmed. It revealed a predominant cluster of protein bands at approximately 25 kDa (Figure 5.1A) that varied in intensity between samples, most of which contained at least two bands at approximately 23 kDa and 25 kDa. Overall urinary protein output, expressed as µg protein per µg creatinine to correct for urine dilution (Beynon & Hurst, 2004), did not differ significantly between males (median 0.21, IQR (interquartile range) 0.14 – 0.28) and females (median 0.15, IQR 0.13 – 0.24; effect of sex: X^2 = 1.85, 1 d.f., p = 0.17; Figure 5.1B). There was also no difference in urinary protein output between samples collected during breeding (spring, autumn: median 0.20, IQR 0.14 – 0.37, n = 8) and non-breeding seasons (summer, winter: 0.18, IQR 0.14 – 0.26, n = 34; X^2 = 0.52, 1 d.f., p = 0.47). The dilution of urine samples, assessed by creatinine concentration did not differ between the sexes (X^2 = 0.002, 1 d.f., p = 0.97) or between seasons (X^2 = 1.03, 1 d.f., p = 0.31). The denotation of animals as displayed in Figure 5.1 will remain consistent within the chapter (M1, M2 etc.).

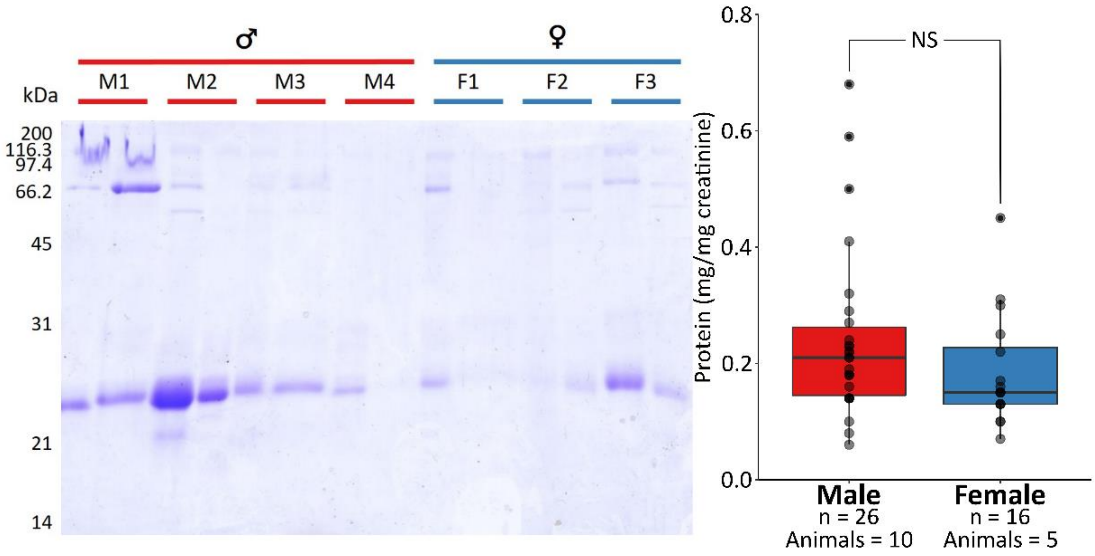


Figure 5.1 | *Investigation into the urinary protein content of T. vulpecula by SDS-PAGE and overall protein content.*
Urine samples from 10 male (M, n=26) and 5 female (F, n=16) T.vulpecula were analysed by SDS-PAGE (A), and the urinary protein content of each sample was determined using the Bradford assay. Protein concentration is displayed as a function of creatinine, to correct for urine dilution (B). No significant differences between male and female urinary output was found (p=0.17)

To obtain further information on protein complexity, urine samples were desalted using Zeba™ Spin Desalting Columns and analysed by ESI-MS. Loading amounts were normalised to 5 pmol protein for male samples and 80 pmol for female samples. In the first instance, mass spectra were deconvoluted over a large range (5 – 100 kDa) to identify the approximate mass. Protein profiles were similar, with signals in a region between 20 and 21 kDa dominating most spectra. Examples from a male and female sample are displayed in Figure 5.2 and Figure 5.3.

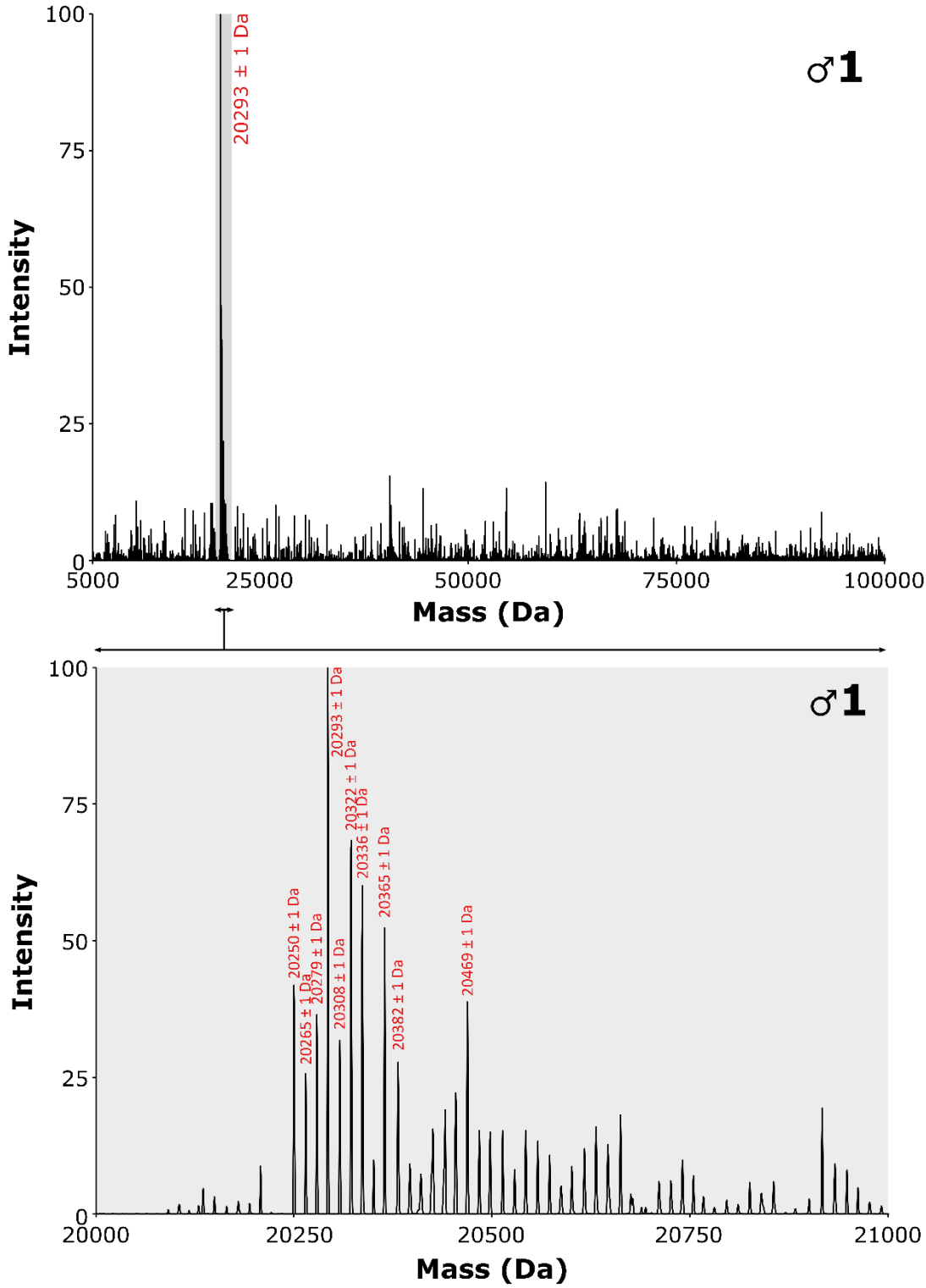


Figure 5.2 | *Investigation into the urinary protein content of T. vulpecula by intact protein analysis.*

Overall protein profiles of *T.vulpecula* urine from 5 male (n=10) and 5 female (n=7) were attained using ESI-MS. Initial wide-range deconvolution revealed the spectrum was dominanted by masses of approximately 20.3 kDa. Spectral deconvolution within a range of 20 – 21 kDa (shaded). An example of a male urine sample is displayed above (sample numbering corresponds to gel lanes in Figure 5.1).

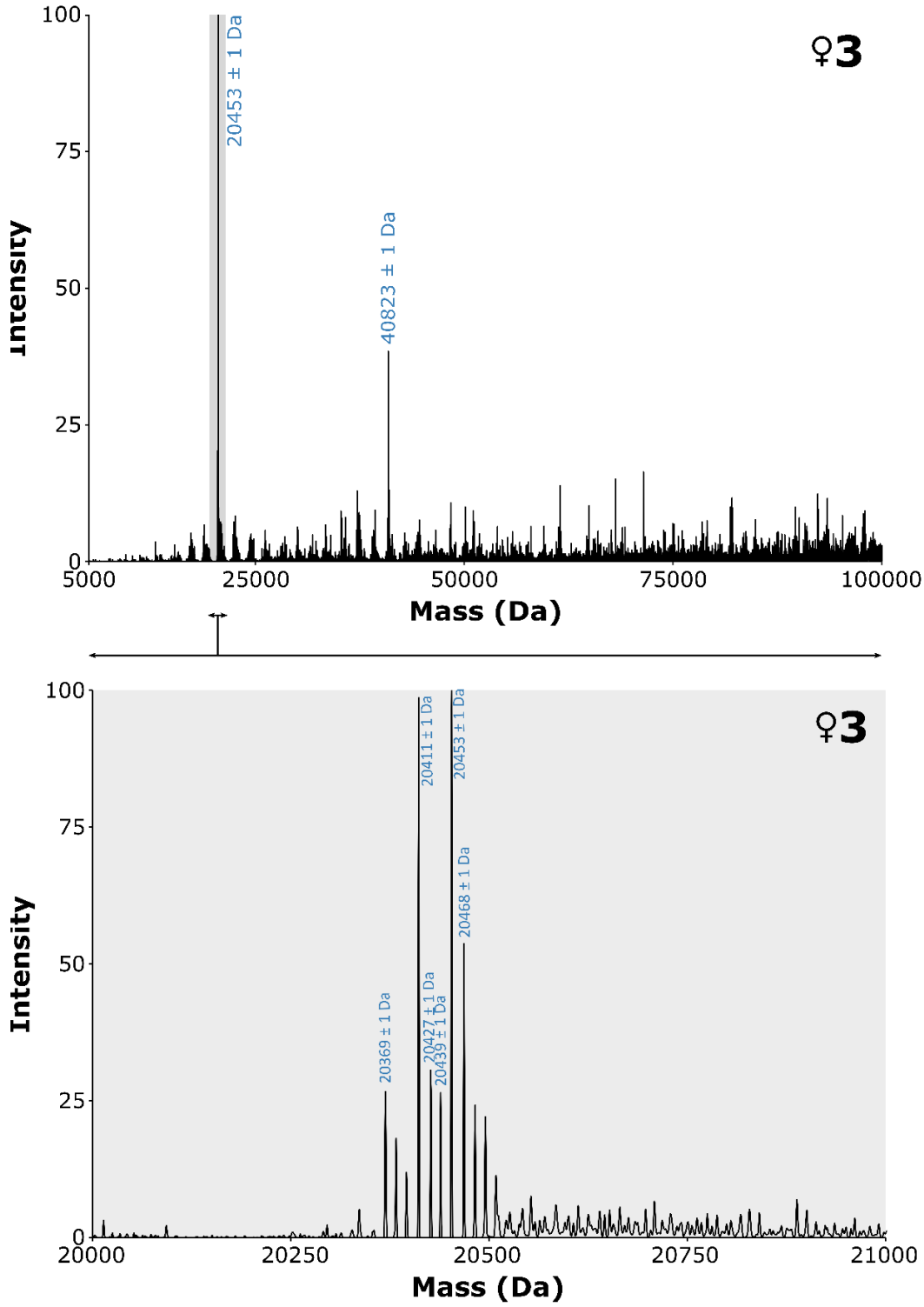


Figure 5.3 | *Initial investigation into the urinary protein content of T. vulpecula by intact protein analysis.*

Overall protein profiles of *T.vulpecula* urine from 5 male (n=10) and 5 female (n=7) were attained using ESI-MS. Initial wide-range deconvolution revealed the spectrum was dominanted by masses of approximately 20.3 kDa. Spectral deconvolution within a range of 20 – 21 kDa (shaded). An example of a female urine sample is displayed above (sample numbering corresponds to gel lanes in Figure 5.1).

The mass was subsequently measured by deconvolution of the spectra to a range of 20 and 21 kDa and a complex protein profile within this range was observed. Abundant peaks were predominantly situated between 20.2 – 20.6 kDa, and a complex profile was recorded, differing in the number and intensity of predicted masses, both between individuals and between replicate samples from the same individual. Despite the diversity of the spectra, the mass differences between peaks displayed consistent patterns. A mass increase of 43 ± 1 Da was frequently observed, in some cases as a predominant series of masses increasing by this value, as in Figure 5.4, Example 1. In other samples, the 43 Da mass increments were interspersed with peaks with mass differences of 14 – 15 Da (Figure 5.4, Example 2).

This is perhaps explained by looking more closely at the original mass spectrum. Whilst the same amount of protein was analysed from each sample, the quality of the spectra are different. Figure 5.5, example 1 is a section of a mass spectrum obtained from analysing a sample in which, when analysed by SDS-PAGE, the 20 kDa protein band was by far the most abundant (Figure 5.1, M2 lane 1). In contrast, the second example is a sample which when visualised by SDS-PAGE appears less dominated by the 20 kDa protein band. Additionally, the overall protein concentration of the former is over double that of the latter (1.4 and 0.6 mg/mL, respectively). Consequently, the 20 kDa protein peaks in example 2 is not only a lower proportion of the total protein injected, but will have been diluted to a lesser degree, and contaminants will therefore cause more interference. The peaks in example 2 are less well defined than those in the first example and the quality of mass prediction is affected accordingly. However, good quality spectra, like that of example 1, still show many other peaks that whilst 'in the grass', are still arguably defined *m/z* peaks that account for the less abundant signals in the de-convoluted spectra.

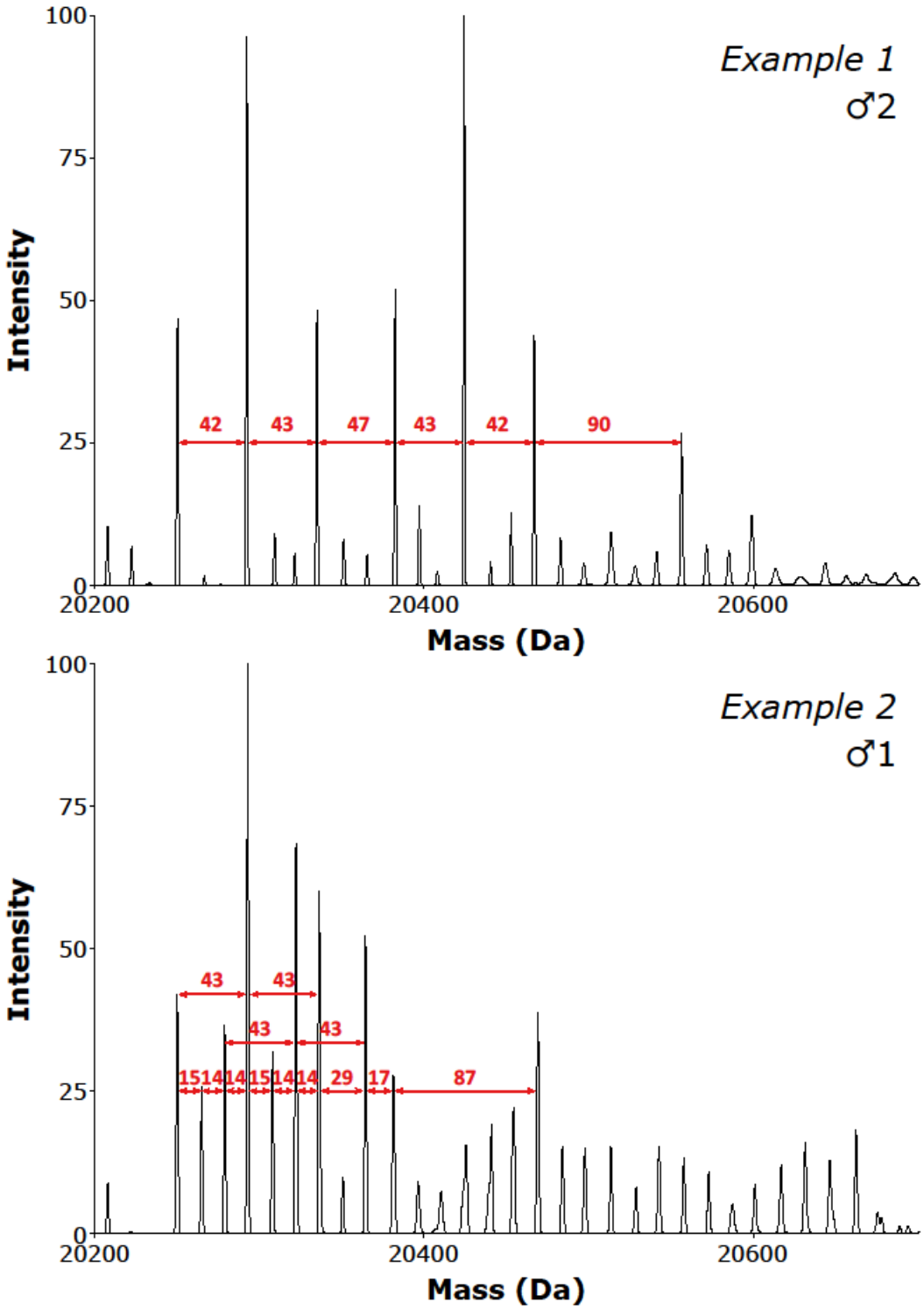


Figure 5.4 | ***Investigation into the urinary protein content of T. vulpecula by intact protein analysis: examples of variation in the deconvoluted mass spectrum.***
Overall protein profiles of T.vulpecula urine samples from 5 males (n=10) and 5 females (n=7) were attained using ESI-MS. A mass series differing by 43 ±1 Da was consistently seen in many samples following spectral deconvolution within a range of 20 – 21 kDa (Example 1 from ♂2). In other samples, a series of smaller shifts was observed within the 43 Da series, consisting of mass increments of 15 or 14 Da (Example 2 from ♂ 1). Spectra were aligned using SpecAlign using the PAFFT method and normalised to TIC.

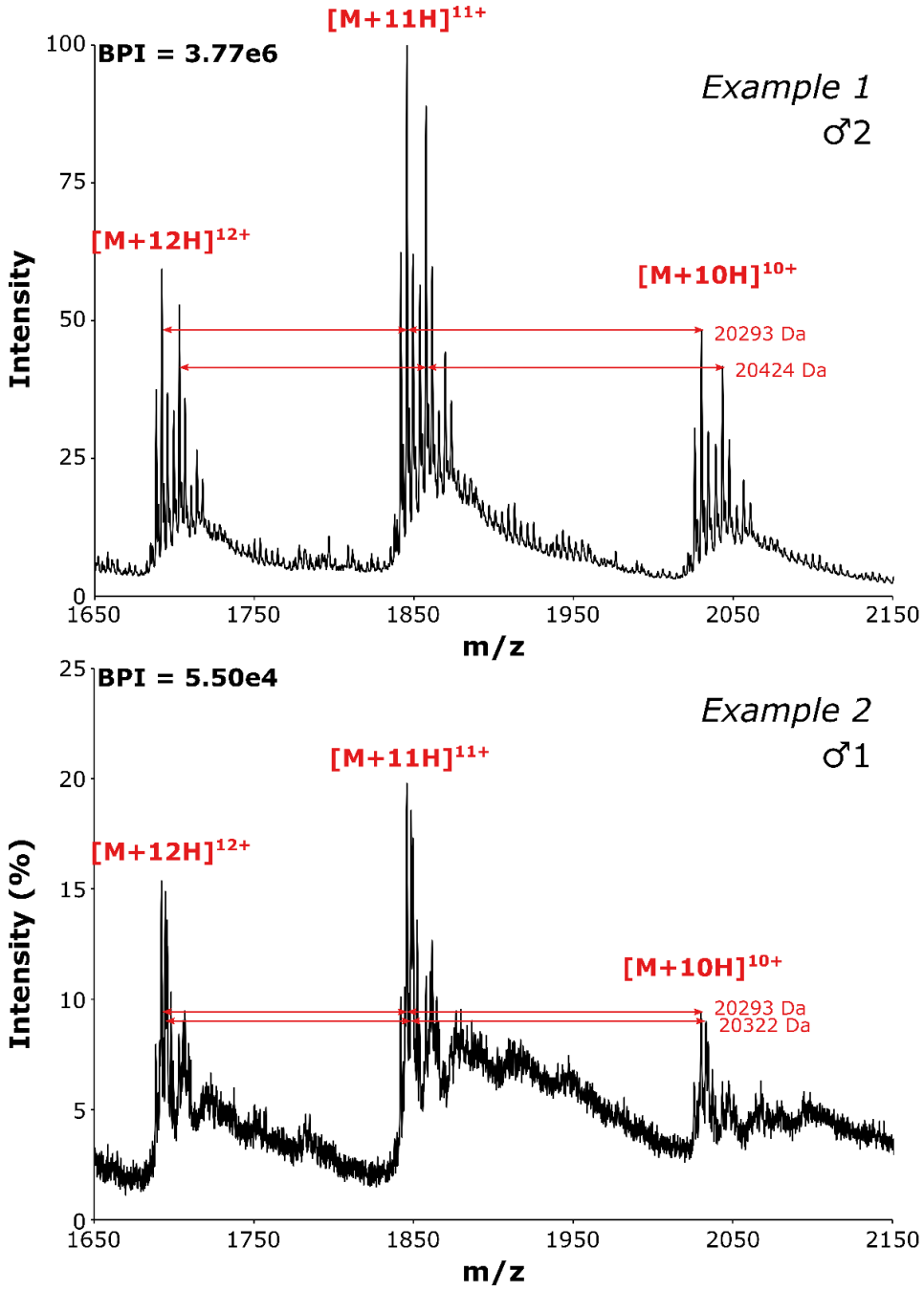


Figure 5.5 | ***Investigation into the urinary protein content of T. vulpecula by intact protein analysis: examples of m/z spectra varying in spectral quality.***
Complexity variation in the deconvoluted spectra can be traced to spectral quality. The mass spectrum of Example 1 has far better resolution than Example 2, despite normalising to overall protein content. Charge states for the two highest intensity series are displayed, with the corresponding overall mass. Spectra were aligned using SpecAlign using the PAFFT method and normalised to TIC.

Heterogeneity is evident even within individual animals. Figure 5.6 is an example of intact mass measurements taken of four samples from the same individual (Male 3), two on different dates in February 2017 and two in July 2017. The July samples had three distinct major masses; 20250, 20293 and 20335 Da. The samples collected in February also had

these, but at much lower abundance, barely visible in the first spectrum which is dominated by a single peak of 20427 Da. The raw spectra for the February samples were of poorer quality and were therefore noisier for the same reasons discussed above. The possibility of a seasonal effect on protein output was considered but due to sampling and time constraints was not pursued.

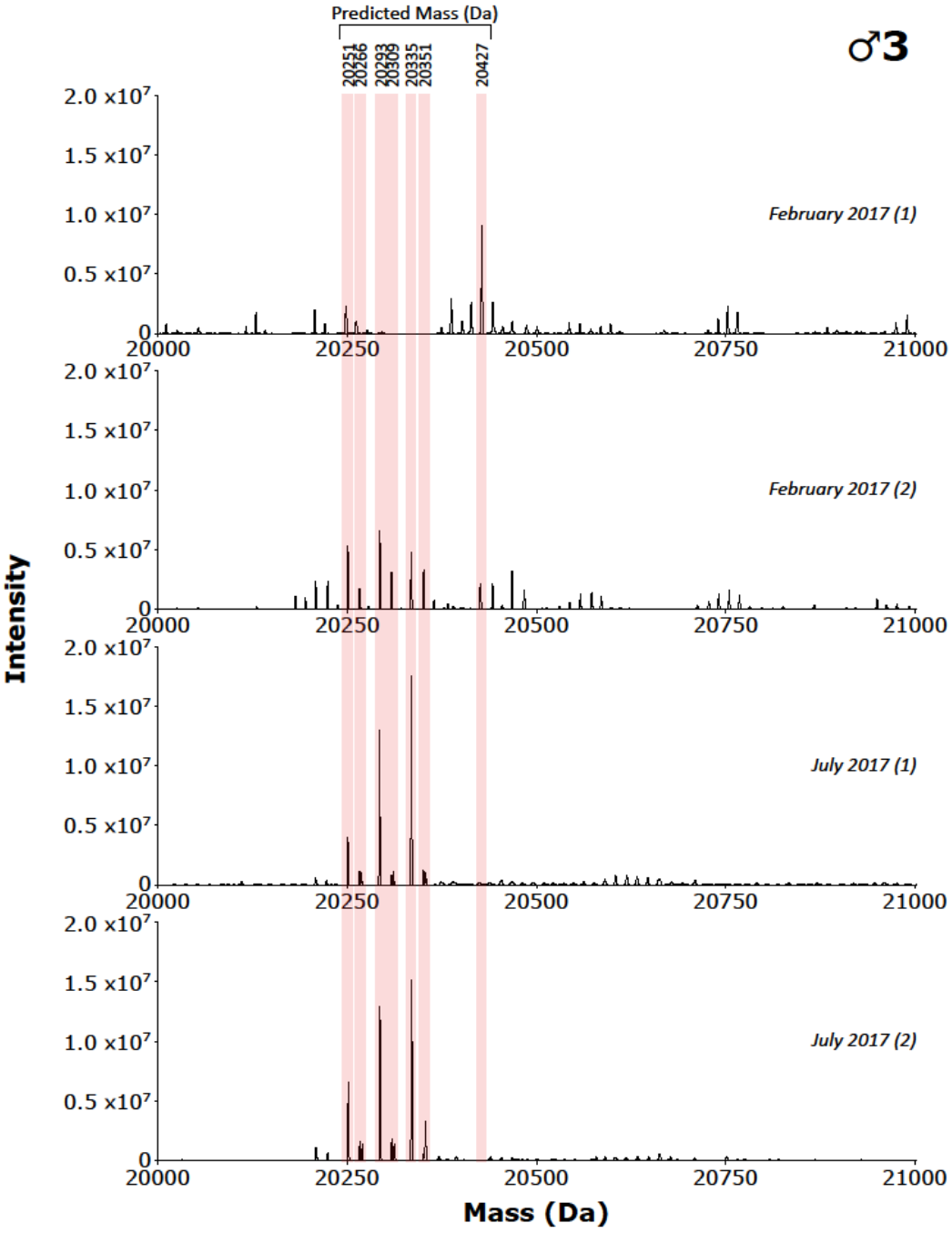


Figure 5.6 | ***Investigation into the protein content of replicate urine samples from the same individual.***
As an example of intra-individual variation, protein profiles from four samples, two collected in February 2017 and two in July 2017, collected from the same male (♂3), were aligned using SpecAlign using the PAFFT method and normalised to TIC.

Deconvoluted spectra from five males (n=10) and five females (n=7) were aligned using the PAFFT algorithm in SpecAlign, and normalised to TIC. Profiles were then displayed as a heatmap and hierarchical clustering of the samples was performed to see if protein profiles were clustered according to individual (Figure 5.7; Figure 5.8).

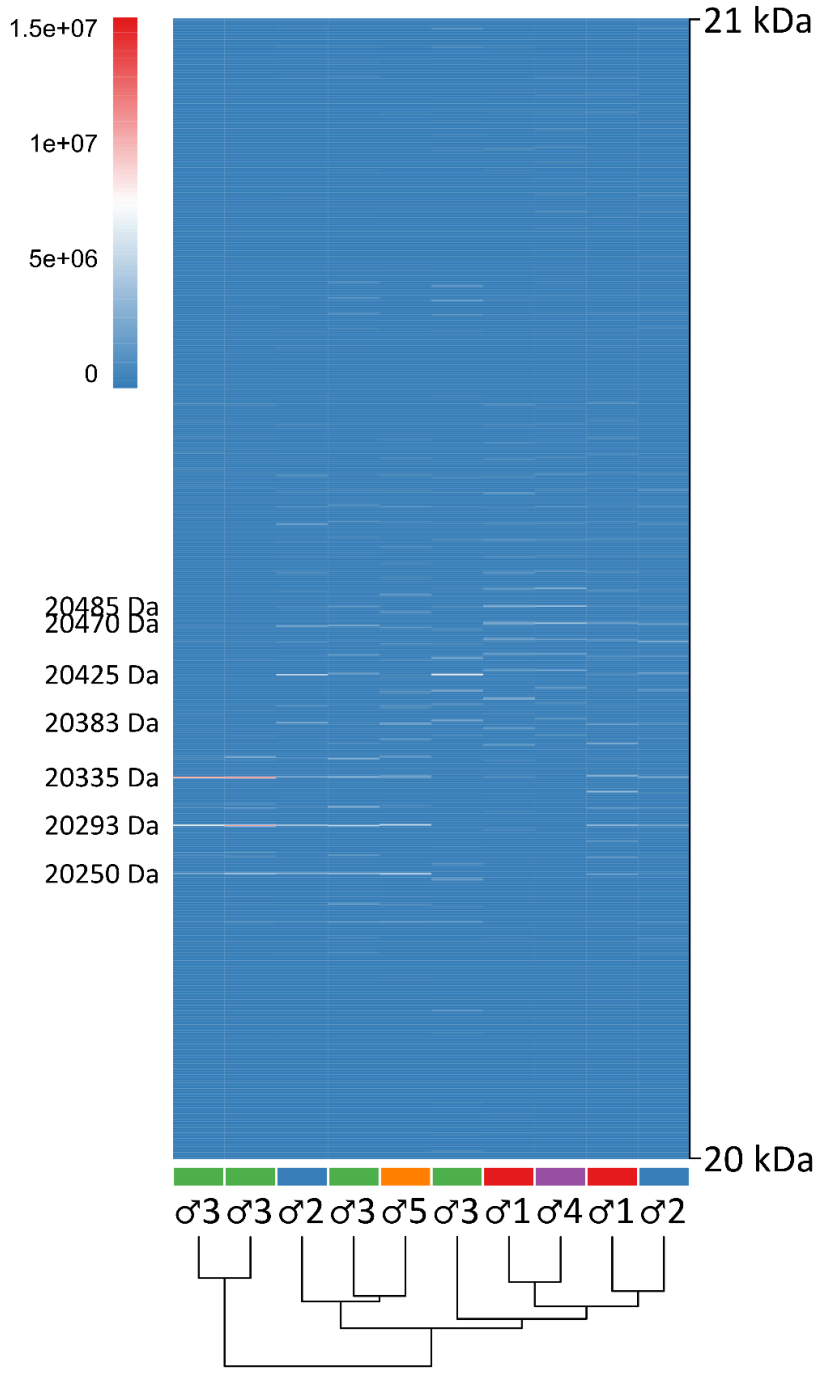


Figure 5.7 | ***Intact protein analysis of ten urine samples from five male brushtail possums.*** Protein profiles between 20and 21 kDa from ten samples collected from the five males were aligne using SpecAlign using the PAFFT method and normalised to TIC. Hierarchical clustering of the samples was used to explore similarities between and within individuals. Mass peaks with intensities higher than 4.02 e^6 are indicated (25% maximum intensity).

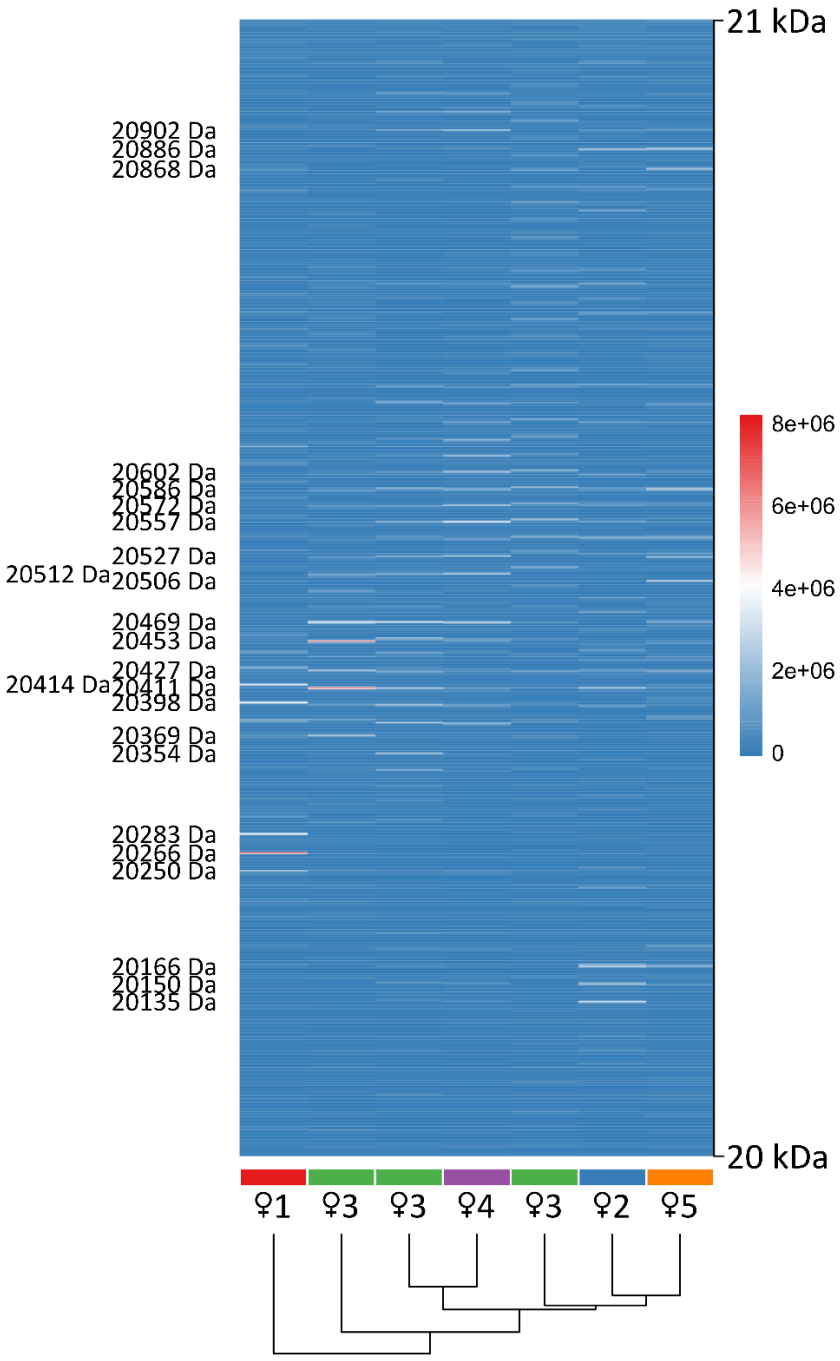


Figure 5.8 | ***Intact protein analysis of ten urine samples from five male brushtail possums.*** Protein profiles between 20and 21 kDa from seven samples collected from five females were aligned using SpecAlign using the PAFFT method and normalised to TIC. Hierarchical clustering of the samples was used to explore similarities between and within individuals. Masses with intensities higher than 2.05 e^6 are indicated (25% maximum intensity).

In neither the male nor female intact mass alignment did the samples appear to cluster by animal. Unfortunately a direct sex comparison could not be performed due to differences required in loading amounts. Overall, male samples appear to be less complex than those from females, with only seven mass peaks above a threshold intensity of 25%. Three

masses were dominant in seven out of ten male samples, 20250, 20293 and 20335 Da. However, this could be skewed by two samples from ♂3, (far left cluster in heatmap), which have a higher base peak intensity than in the other samples. As explored earlier, their 'cleaner' profile could be explained by a higher overall protein concentration, required a higher fold dilution for analysis and therefore a reduction of contaminants. They were also part of a separate group of samples collected and transported at a different time, and although every effort was taken to ensure consistency in sample collection and handling, there is a possibility that small differences could play a role in changes to the protein profile, especially if this complexity is due to post-translational modifications. Of these three dominant masses, only one is observed above the threshold intensity in females (20250 Da). The protein profiles of females appear to be more complex, with more peaks above 25% intensity across a larger mass range, from 20135 to 20902 Da. Nonetheless, the same incremental patterns of 43 ± 1 Da, 14 ± 1 and 15 ± 1 Da occur frequently.

Intact mass analysis supports a complex protein profile, differing between both individuals and between samples taken from the same individuals at different times. However, the recurring mass increments of 43 ± 1 Da and 14-15 ± 1 Da are not consistent with complexity due solely to protein heterogeneity. Whilst these incremental masses could be attributable to single amino acid mutations, the frequency at which they are observed is more likely an indication of post-translational modifications. Acetylation (+42 Da), oxidation (+16 Da) and methylation (+14 Da) are all common PTMs that could explain these differences within the constraints of instrument error.

To investigate heterogeneity of the urine samples, a peptide mass fingerprint (PMF) was generated from in-gel lys-C digestion of excised gel bands followed by MALDI-ToF mass spectrometry. Despite the heterogeneity of the intact mass profiles, the PMFs of the protein band across male individuals were simple and identical. PMFs of the 23 kDa and 25 kDa bands from SDS-PAGE were also consistent (Figure 5.9; Figure 5.10). Of note, not only were the masses detected homogeneous, but also the relative abundance patterns of peaks across spectra. Female samples also shared many of the peptide peaks, but were not as uniform. This is likely due to the lower quality of spectra caused by the lower protein abundance in female samples, and the higher abundant spectra gained (♀3, 23 kDa; ♀4, 25 kDa; ♀5, 23 kDa) all show strong similarity to the male spectra. If the peptide masses of the digested protein were uniform, in both m/z and relative abundance, it was highly likely that the heterogeneity in the intact mass was caused by a post-translational modification, rather than protein-driven heterogeneity.

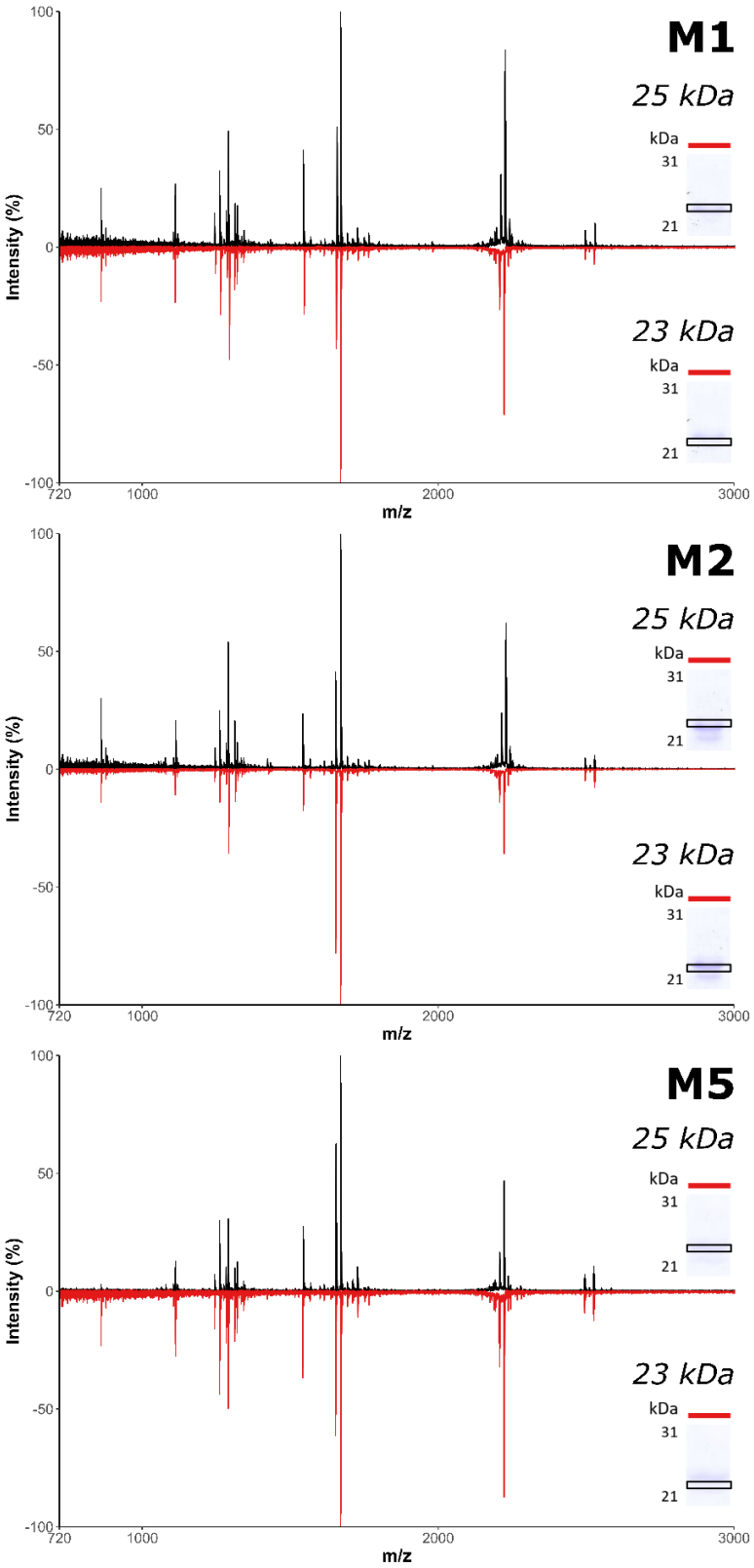


Figure 5.9 | *MALDI-ToF analysis of in-gel digested protein separated by SDS-PAGE and resolving at approximately 23-25 kDa from male brushtail possum urine.*
Both SDS-PAGE gel bands from the doublet at 21 kDa were excised from three male (A) and three female (B) urine samples. Gel plugs were subject to in-gel digestion with Lys-C and analysed by MALDI-ToF with an m/z range of 720-5000. PMFs from slower-migrating gel bands (black) are displayed above the axis and those from the corresponding faster-running bands are displayed below the axis (red).

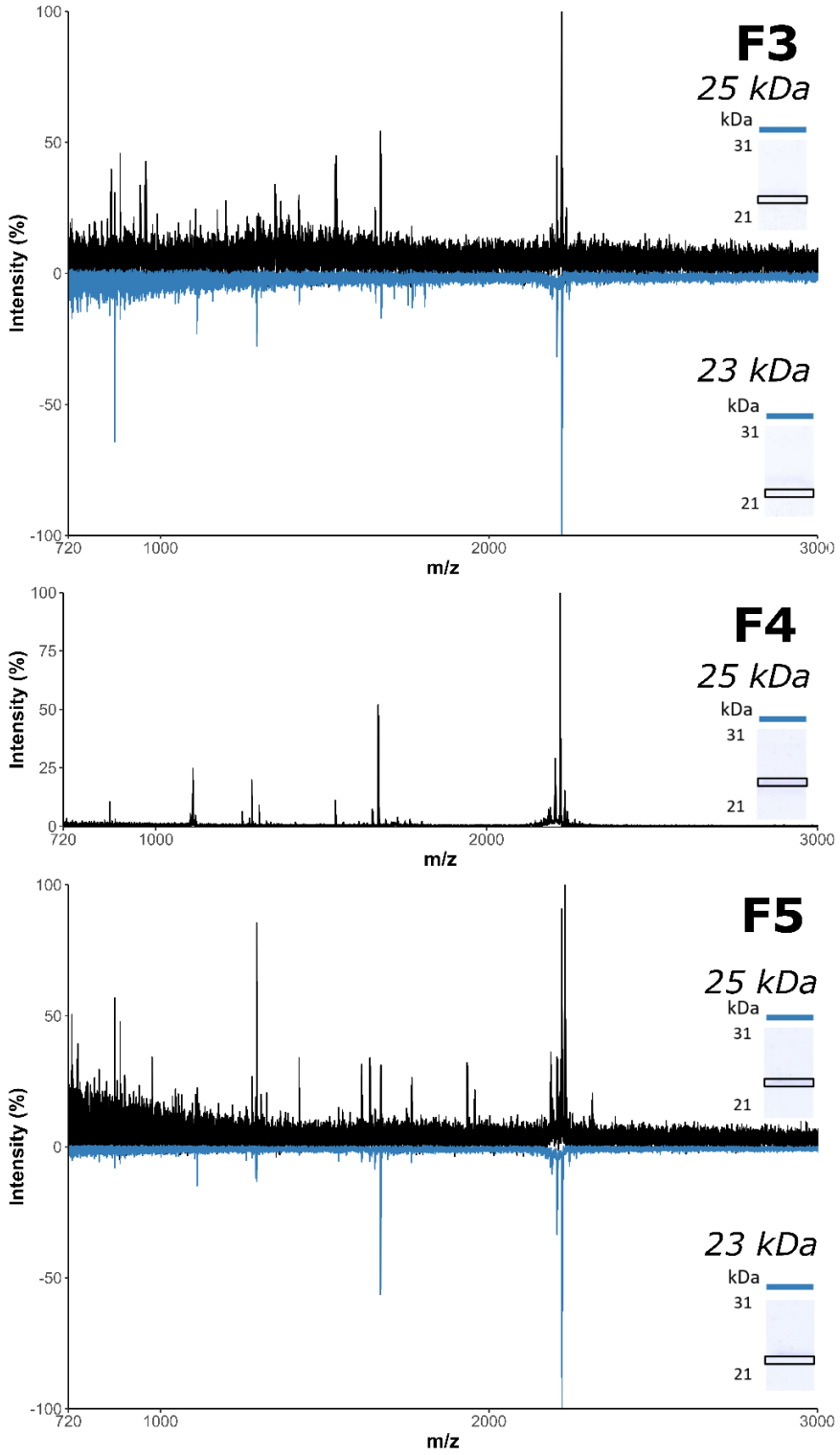


Figure 5.10 | *MALDI-ToF analysis of in-gel digested protein separated by SDS-PAGE and resolving at approximately 23-25 kDa from female brushtail possum urine.*
Both SDS-PAGE gel bands from the suspected doublet at 21 kDa were excised from three male (A) and three female (B) urine samples. Gel plugs were subject to in-gel digestion with Lys-C and analysed by MALDI-ToF with an m/z range of 720-5000 Da. PMFs from slower-migrating gel bands (black) are displayed above the axis and those from the corresponding faster-running bands are displayed below the axis (blue). A PMF for the faster migrating gel band for ♀4 could not be generated.

5.4.2 Protein identification

Preliminary exploration of the protein content of brushtail possum urine exposed strong evidence for a protein, or proteins, of approximately 20 kDa in male and female urine. Whilst intact mass analysis revealed complex, heterogeneous profiles in this region, peptide mass fingerprinting of the digested protein bands suggested a simple, identical protein in both males and females.

To gain a deeper understanding of the protein content, whole urine from seven males and five females were individually digested in-solution with trypsin and analysed by LC-MS/MS. Tandem mass spectrometry data were searched against databases comprising all mammalian proteins in SwissProt, or all protein sequences inferred from an unannotated genome sequence of *M.domestica*. No significant hits were reported that corresponded to the 25 kDa gel band, and the sequences of many of the most abundant peptides generated *de novo* were unidentified. To supplement protein discovery, peptide mass fingerprints (PMF) generated from MALDI-ToF mass spectrometry were used in combination with a list of highly abundant, confident peptide sequences generated *de novo* in the PEAKS™ package (Figure 5.11). The 21 kDa protein band from four male (n=7) and one female (n=2) *T.vulpecula* urine samples were excised and digested with trypsin. Peptides were extracted and analysed by MALDI-ToF. Masses commonly detected in the PMFs (Figure 5.12) were used to retrieve corresponding peptides sequenced by PEAKS™ from the LC-MS/MS data (Table 5.1). These were searched individually against non-redundant mammalian protein sequences in BLAST. The most consistent highly-scoring sequence was the trichosurin-like protein from *Monodelphis domestica* (NCBI Reference Sequence: XP_007475413.1 or UniProt accession: F7F0X2). Nine selected peptides covered 38% of the homologous protein after the signal peptide was removed by Signal P Server (Petersen *et al.*, 2011), at 55-80% identity (Figure 5.13).

PEAKS SPIDER searches are built with the intention of identifying mutations and reducing *de novo* sequencing error. However, all nine sequenced peptides contained at least two non-consecutive mutations in respect to the *M.domestica* protein and consequently these peptides were not identified by PEAKS™.

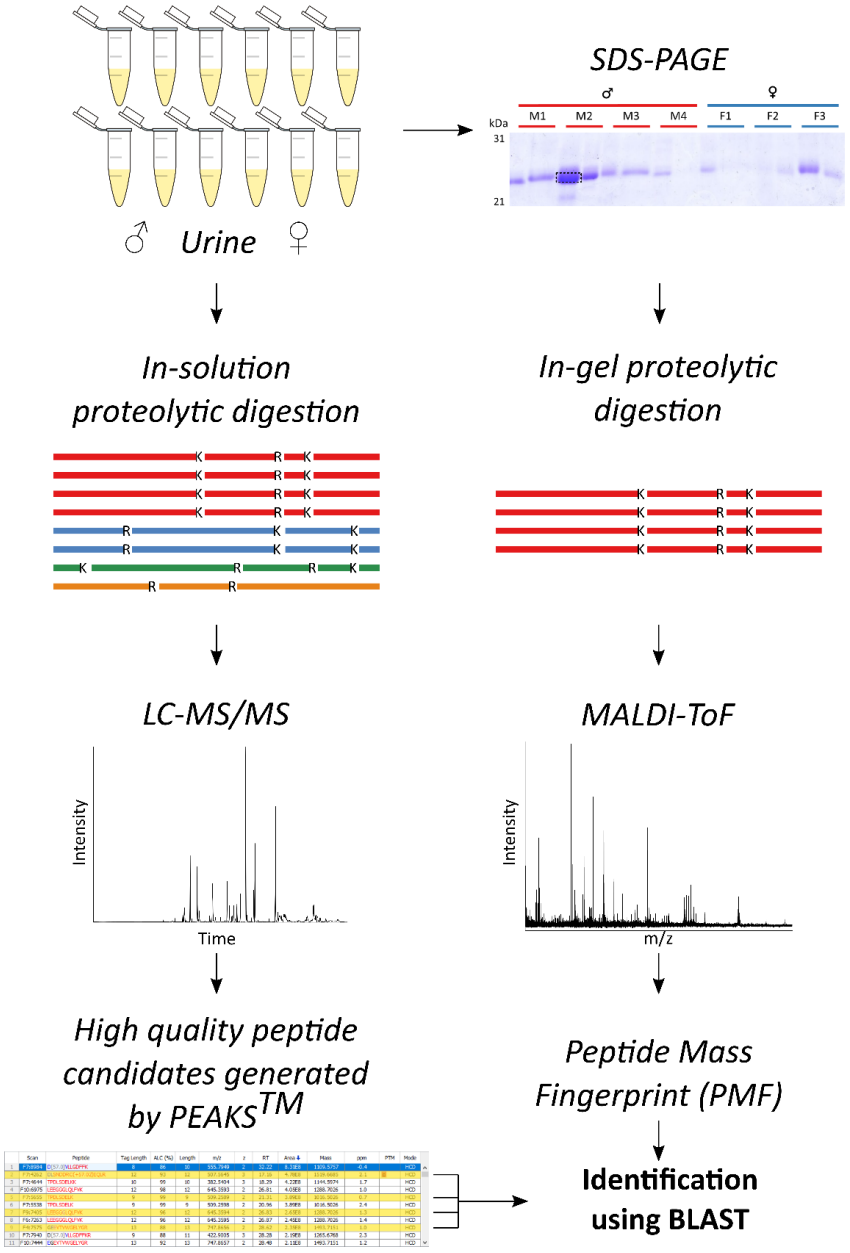


Figure 5.11 | *Protein identification.*
Two parallel proteomic approaches were used in the identification of the abundant *T.vulpecula* protein. Urinary proteins in samples from seven male and five female *T.vulpecula* were digested with trypsin and analysed by LC-MS/MS. Peptide data were analysed in PEAKS™, from which MS/MS fragment ion data were used to predict sequences *de novo*. Concurrently, urine samples were analysed by SDS-PAGE, and the resulting 21 kDa gel bands from 4 male (n=7) and 1 female samples (n=2) were excised, and the isolated protein digested with trypsin. MALDI-ToF mass spectrometry analysis of these digests generated PMFs from which peptide masses were matched to high-quality sequences generated *de novo* from the whole urine digest. Each peptide match was searched against non-redundant mammalian sequences in the NCBI database using BLAST.

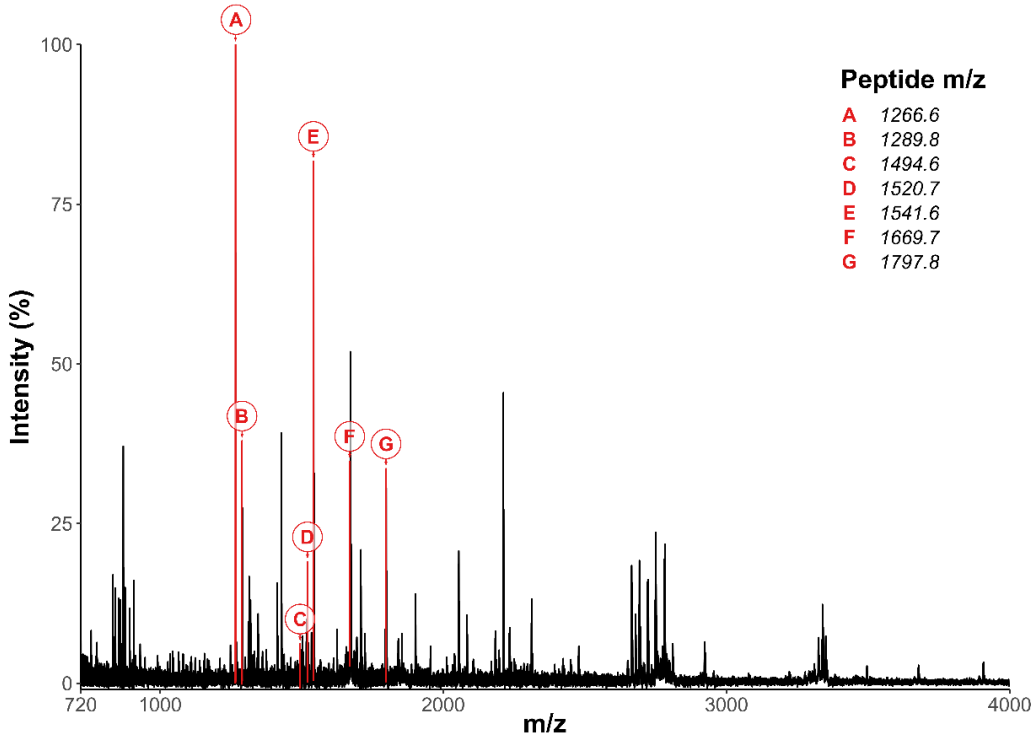


Figure 5.12 | *Protein Identification.*
Protein-containing bands were removed from SDS-PAGE gels and subject to in-gel protease digestion. Digested protein from 4 males (n=7) and 1 female (n=2) *T.vulpecula* urine samples were analysed by MALDI-ToF MS. PMFs for all 9 samples were similar and can be found in Supplementary S5.2. An example is displayed above. Peaks corresponding to *de novo* sequences generated from PEAKS™ are labelled in red.

Table 5.1 | *Protein identification.*
High-scoring, abundant peptides generated in PEAKS™ that match PMF masses. For each, the predicted sequence is stated with the best average local confidence (ALC) score, mass, charge, retention time (t_R) and area for each charge state (2+ and 3+).

#	PMF *m/z*	Predicted Peptide	ALC (%)	Mass	$[M+nH]^{n+}$	n	t_R	Area
A	**1266.6**	*DGVJJGDFFKR*	97	1265.670	633.8463	2	28.31	2.41E+07
			89		422.9004	3	28.29	1.65E+07
B	**1289.8**	*JEEGGGJQJFVK*	99	1288.703	645.3592	2	26.93	5.20E+07
			76		430.5747	3	26.90	2.75E+07
C	**1494.6**	*GEEVTVWGEJYGR*	92	1493.715	747.8657	2	28.48	2.11E+08
			91		498.9128	3	28.61	3.08E+06
D	**1520.7**	*DLSNDDRCEQLR*	98	1519.669	760.8426	2	17.14	1.09E+07
			94		507.5646	3	17.25	1.34E+08
E	**1541.6**	*FEELCENFGLRK / KFEELCENFGLR*	99	1540.734	771.3759	2	23.59	6.43E+07
			93		514.5864	3	22.45	4.25E+07
F	**1669.7**	*KKFEELCENFGLR / KFEELCENFGLRK*	98	1668.829	835.4229	2	21.56	8.50E+06
			97		557.2833	3	21.12	1.15E+07
G	**1797.8**	*KKFEELCENFGLRK*	96	1796.924	599.9827	3	19.69	4.17E+06
			96		899.4705	2	19.70	1.28E+06

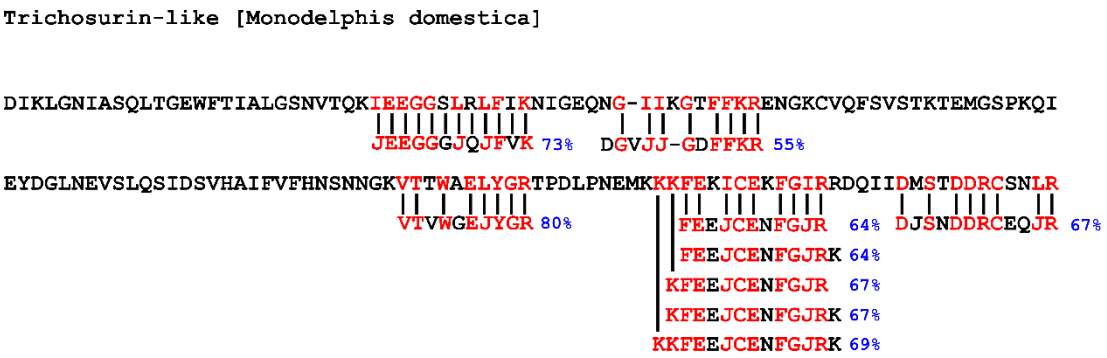


Figure 5.13 | *Protein Identification*.
Abundant peptide masses generated from in-gel digestion were used to identify corresponding *de novo* sequences in PEAKS, which when searched using BLAST against all non-redundant mammalian protein sequences inferred homology to the trichosurin-like protein from *Monodelphis domestica* (NCBI Reference Sequence: XP_007475413.1 or UniProt accession: F7F0X2). The *M.domestica* sequence was aligned with identified peptides. Exact amino acid matches are highlighted in red, and sequence identity (%) is denoted in blue. The letter 'J' is used to represent leucine or isoleucine, and was specified to match either in the homologous sequence.

The trichosurin-like protein from *M.domestica*, 17.5 kDa with signal peptide removed, is a lipocalin. It contains the conserved lipocalin motif [G-X-W], and contains three conserved cysteine residues homologous with MUPs. Its name derives from trichosurin, a milk whey lipocalin discovered in *T.vulpecula*, of which both protein sequence and structure have been experimentally solved (Watson *et al.*, 2007). It is so far the only experimentally confirmed marsupial lipocalin sequence.

To assist protein sequencing, the trichosurin-like protein from *M.domestica* was searched against all non-redundant mammalian protein sequences in the NCBI database using BLAST, and the 25 top-scoring hits (manually excluding truncated protein sequences) was used to construct a multiple sequence alignment in MUSCLE (Edgar, 2004) which was then annotated in JalView (Waterhouse *et al*., 2009) (Figure 5.14). This aimed to identify highly conserved regions of these protein sequences, and was subsequently used as a scaffold to support protein sequencing. The multiple sequence alignment consisted of a mixture of lipocalin sequences. Other marsupial sequences from *Phascolarcturus cinereus* (koala) and *Sarcophilus harisii* (Tasmanian devil) were identified in addition to trichosurin. All were denoted as having homology to trichosurin, MUPs, epididymal-specific lipocalins and allergen lipocalins. Each of these lipocalin families was also identified from other species, however no specific group scored consistently highly enough above others to suggest the trichosurin-like protein belonged to one of these groups; sequence identity only ranged from 36-57 %.


Predicted sequence T.vulpecula
A Trichosurin T.vulpecula
B Trichosurin-like S.harrisii
C Minor allergen Can f 2-like S.harrisii
D Trichosurin-like isoform X1 S.harrisii
E Trichosurin-like isoform X2 S.harrisii
F Trichosurin-like isoform X1 M.domestica
G Trichosurin-like isoform X2 M.domestica
H Trichosurin-like M.domestica
I Major urinary protein-like P.cinereus
J Major urinary protein-like isoform X2 P.cinereus
K Trichosurin-like isoform X1 P.cinereus
L Epididymal-specific lipocalin-9 P.cinereus
M Allergen Fel d 4-like H.armiger
N Allergen Fel d 4-like D.ordii
O Major urinary protein 4-like C.cristata
P Alpha-2u globulin R.norvegicus
Q Odorant Binding Protein 3 R.norvegicus
R Major urinary protein-like R.norvegicus
S Major urinary protein isoform X1 R.norvegicus
T Allergen Fel d 4-like isoform X1 D.rotundus
U Allergen Fel d 4-like isoform X2 D.rotundus
V Major allergen Equ c 1-like G.variegatus
W Epididymal-specific lipocalin-9 M.p.furo
X Major urinary protein 20-like I.tridecemlineatus
Y Major urinary protein 20-like M.m.marmota

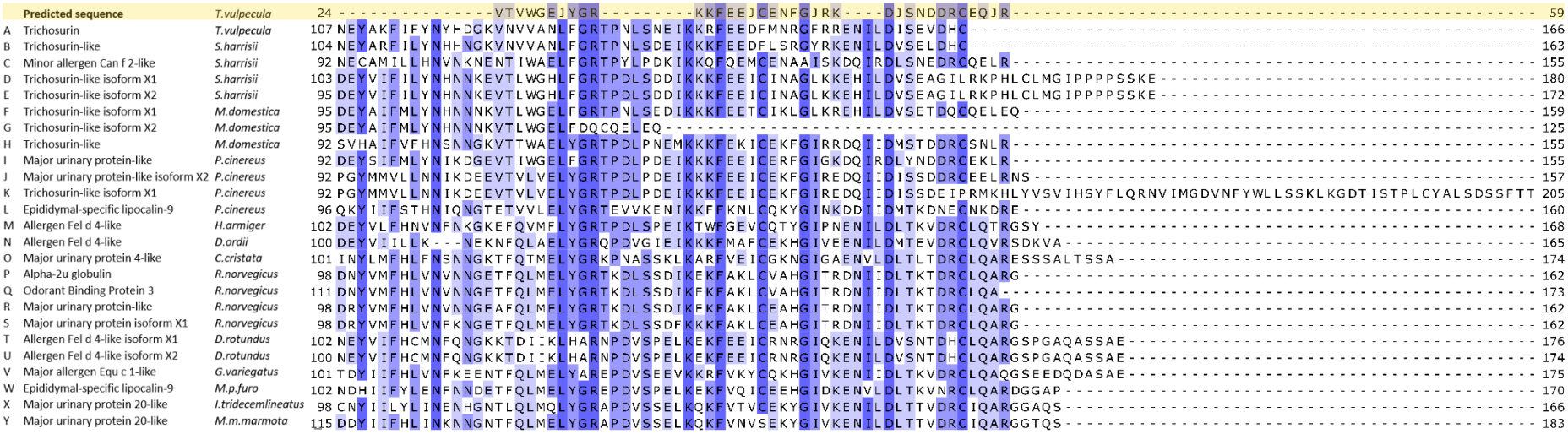


Figure 5.14 | *Protein Identification*.
Homologous sequences to the trichosurin-like protein from *Monodelphis domestica* (NCBI Reference Sequence: XP_007475413.1 or UniProt accession: F7F0X2) were identified using BLAST. The 25 top-scoring mammalian protein sequences in the NCBI database were aligned using MUSCLE (Edgar, 2004), in addition to the newly identified peptides, and annotated in JalView (Waterhouse *et al*., 2009). Residues are coloured by conservation (%). The sequence highlighted in yellow is constructed of new urinary *T.vulpecula* peptides.

5.4.3 Protein Purification

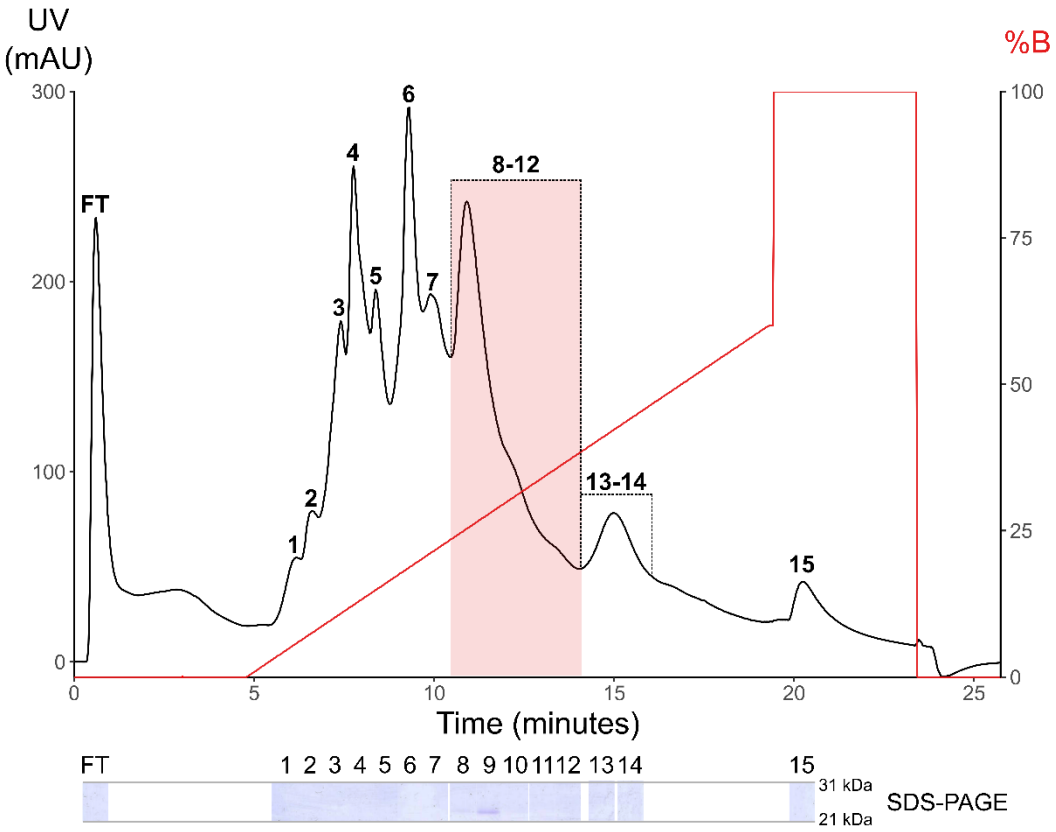


Figure 5.15 | *Protein purification*.
Proteins from pooled male *T.vulpecula* urine were loaded onto a Resource Q™ 1 mL column and eluted off a gradient from 0 – 0.6 M NaCl in 20 mM Tris HCl pH 8.5 over 30 column volumes. Fractions were manually collected (numbered FT, flow through, to 15) according to a UV absorbance of 280 nm. Fractions were subject to SDS-PAGE analysis and protein concentration was assessed using the Bradford method. Fractions of the protein-containing peak (red box) were pooled for further analysis. Brightness and contrast of SDS-PAGE image were adjusted for clarity.

Urinary proteins were further resolved by anion exchange chromatography (Figure 5.15). Male urine samples were pooled and dialysed against 20 mM Tris-HCl, pH 8.5, after which protein (350 µg) was separated by anion exchange chromatography using a 1 mL Resource Q™ column with a 0 to 0.6 M NaCl gradient over 30 column volumes. Eluent corresponding to protein peaks with high absorbance at 280 nm were collected manually. SDS-PAGE was used to identify the elution position of the 20 kDa protein(s). These fractions were pooled and the semi-purified protein(s) were concentrated using Strataclean™ resin prior to proteolysis and tandem mass spectrometry. Intact mass analysis confirmed that the mass profile previously observed in pooled whole urine co-eluted at the same NaCl concentration in the gradient (Figure 5.16).

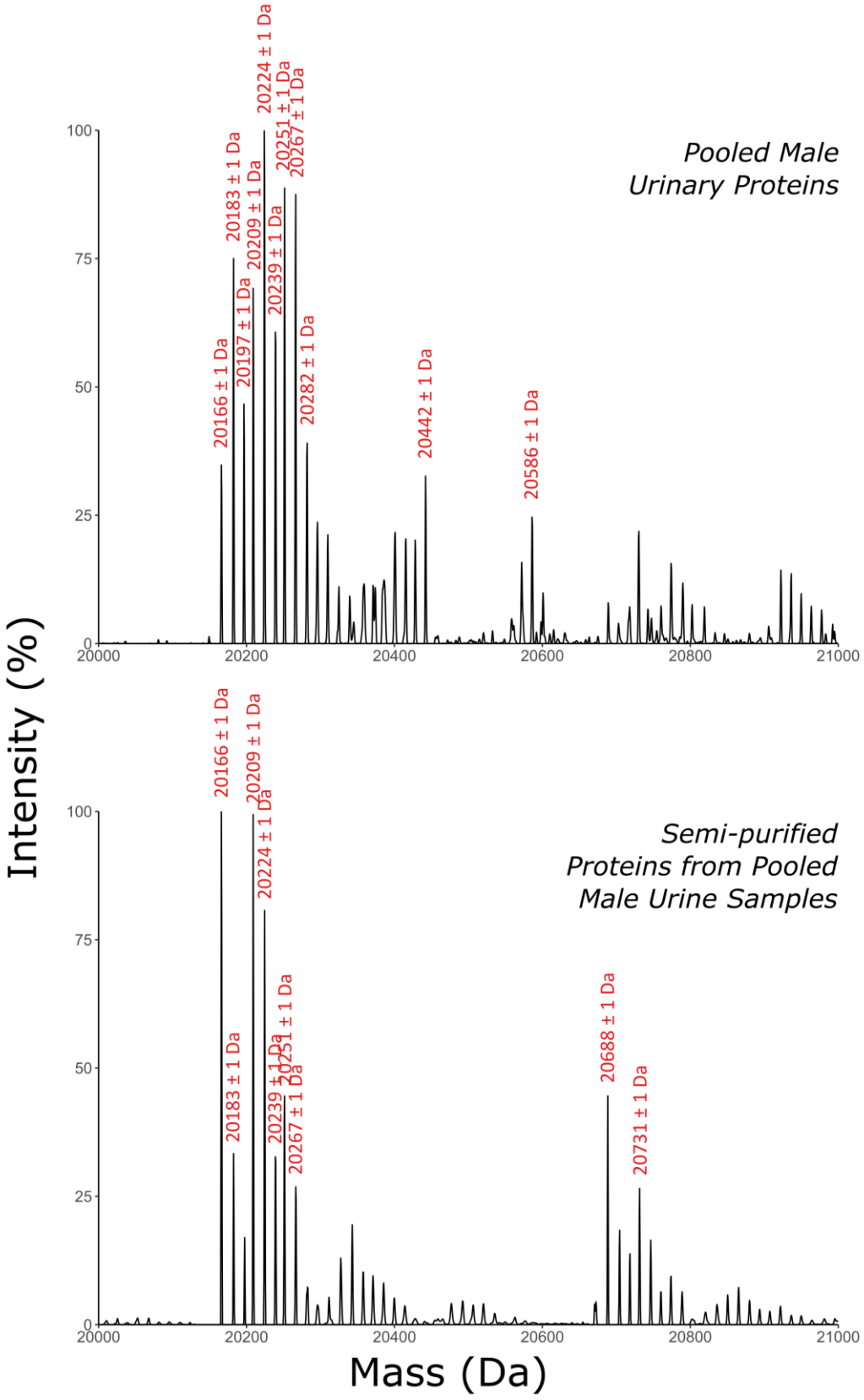
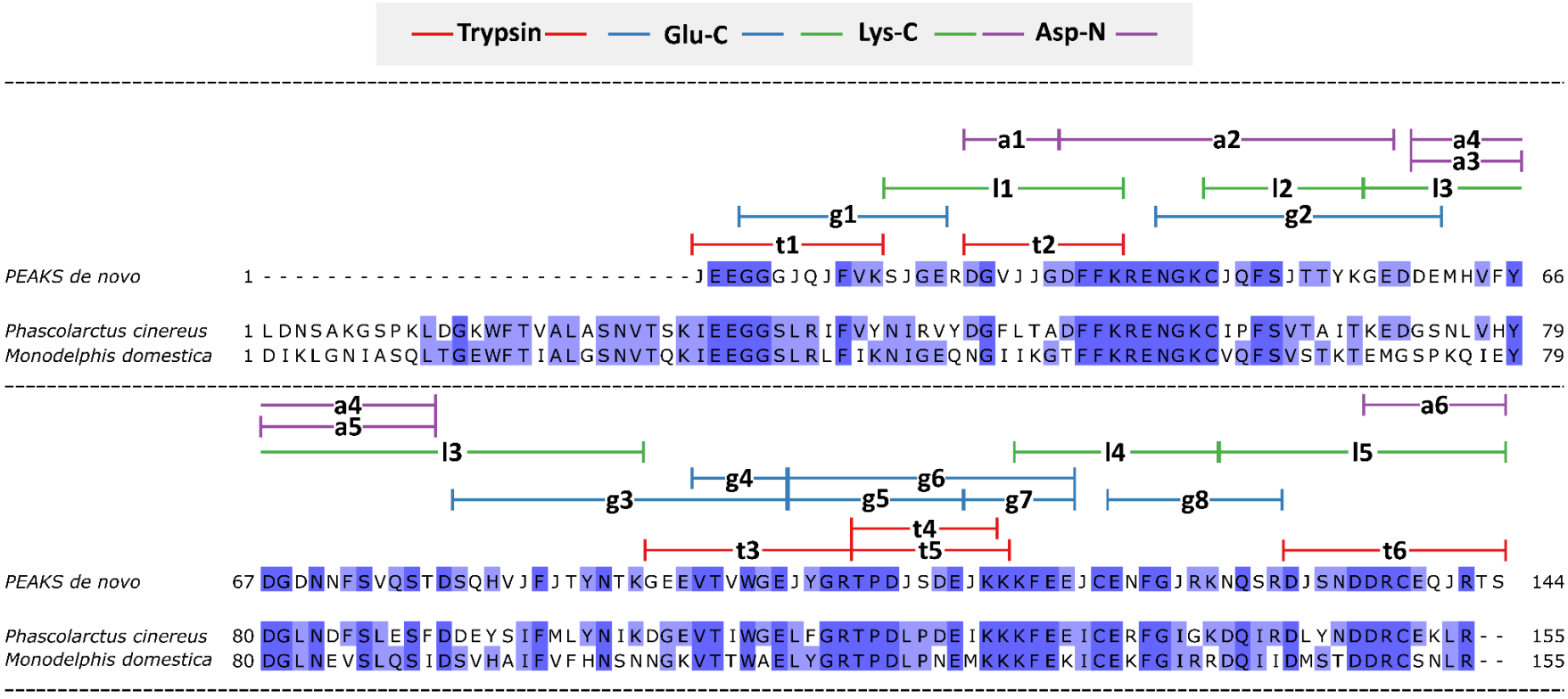


Figure 5.16 | *Protein purification*.
Protein purified using anion exchange chromatography was concentrated and desalted prior to ESI-MS analysis. Mass spectra were obtained for the pooled male urine (below) and compared to the original pooled urine sample (above).

5.4.4 Protein sequencing

The protein isolated from anion exchange chromatography (Fraction 9) was digested independently with either trypsin, endopeptidase Lys-C, endopeptidase Glu-C or endopeptidase Asp-N to generate sets of peptides to be analysed by LC-MS/MS. The resulting LC-MS/MS data were analysed in PEAKS^{TM} and high-quality *de novo* sequences were used to construct the protein sequence with manual checking. Peptide sequence data were iteratively searched against the developing protein sequences in PEAKS to assist sequencing *de novo*, and the multiple sequence alignment of homologous proteins (Figure 5.14) was used as a guide for conserved residue calls. The protein was confidently sequenced, apart from isobaric leucine and isoleucine (denoted 'J') (Figure 5.17; fragment ion spectra available in supplementary material) and a re-search of the final sequence generated a peptide map with good coverage (Figure 5.18). Peptide sequence coverage was so high that no region was confirmed by fewer than two overlapping peptides. No consistent mutation calls were discovered in PEAKS^{TM} SPIDER searching, suggesting limited heterogeneity with respect to possible protein isoforms. However, a gap of approximately 27 amino acids near the N-terminus differentiated the sequenced protein from aligned homologous proteins, within which the fully conserved lipocalin motif would G-X-W reside. Whilst this could be the result of a genuine deletion, the calculated average mass of the sequenced protein was 14948 Da, whereas the measured average masses were 20252 Da, 20294 Da or 20335 Da. Thus, the protein sequence obtained was smaller than observed masses by approximately 5 kDa. The average molecular weight of an amino acid is 110 Da, so a missing portion of 27 amino acids would be approximately 3 kDa. Consequently, the missing N-terminus is unlikely to be the reason for this mass deficit. Considering the implication for a post-translational modification from the incremental masses observed in intact mass data, combined with strong evidence for limited protein-driven heterogeneity, a glycosylation modification was hypothesised.


Trypsin
Glu-C
Lys-C
Asp-N
a1
a2
a4
a3
l1
l2
l3
g1
g2
t1
t2
PEAKS de novo 1 - - -

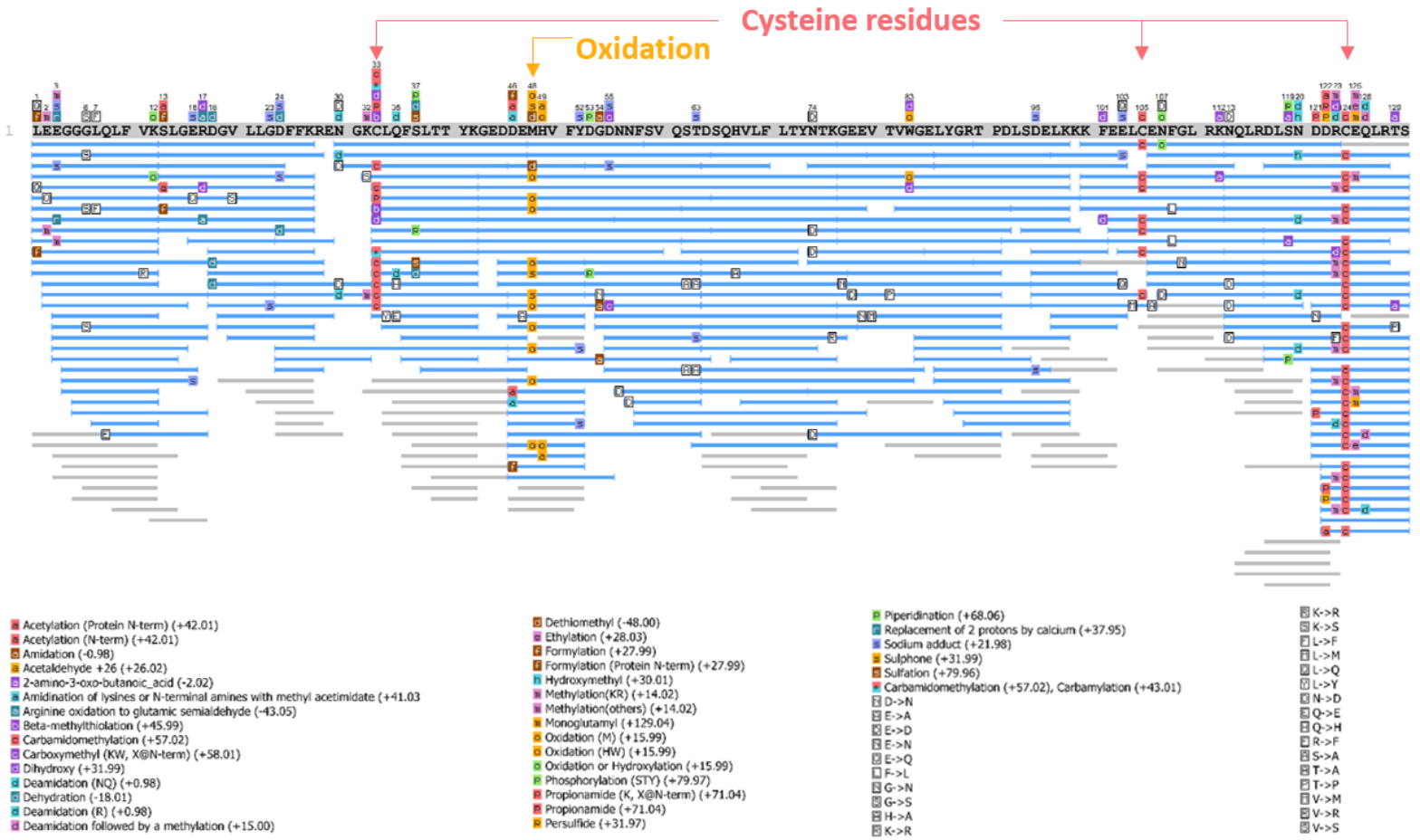


Figure 5.18 | *Sequencing de novo*.
Peptide data were re-searched against the novel protein sequence. Complete coverage was observed from PEAKS^{TM} SPIDER searches, including peptide spectrum matches (blue) and *de novo* tags (grey). No confident mutations (white boxes) were found after manual inspection. Peptides with a -10lgP score ≥ 20 were accepted. The minimum PTM A Score was -10lgP ≥ 20 and mutations with greater ion intensity of 5% are displayed.

5.4.5 Deglycosylation

To investigate the possibility of glycosylation, pooled male urine was incubated with the enzyme PNGase F, which removes *N*-linked glycans (Maley *et al*., 1989). PNGase F has broad specificity, cleaving most *N*-linked glycan moieties at the *N*-glycosidic bond. The two exceptions are when the *N*-linked glycan is attached on the N-terminus (Fan and Lee, 1997) or if the saccharide immediately attached to the core N-acetylglucosamine (GlcNAc) is an α(1,3)-fucose (Tretter, Altmann and März, 1991). The resulting reaction mixture was analysed by SDS-PAGE and a mobility shift was observed, consistent with a deglycosylation reaction (Figure 5.19). The mobility shift also resulted in a shift from one strongly staining gel band resolving at approximately 25 kDa, with a less distinct, fainter band above, to two distinct gel bands resolving at approximately 18 and 21 kDa after deglycosylation. The following section details how the deglycosylation protocol was adapted for both in-solution digestion and LC-MS/MS analysis for completion of sequencing, and ESI-MS, for intact mass analysis.

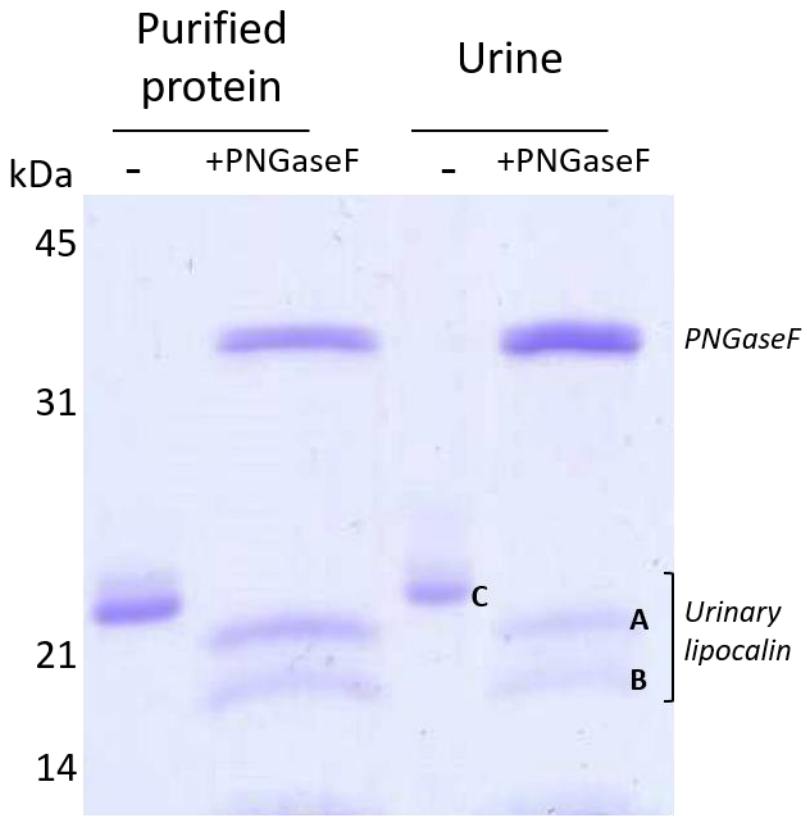


Figure 5.19 | *Deglycosylation of pooled male urinary protein from T.vulpecula*.
Deglycosylation of the newly discovered protein using PNGase F under denaturing conditions, was analysed by SDS-PAGE. Brightness and contrast of image was adjusted for clarity. A and B are bands excised from the deglycosylated sample for in-gel digestion and LC-MS/MS analysis, and C indicates the band containing the intact glycoprotein.

The presence of an *N*-linked glycan attached to the possum protein was determined initially using PNGase F (New England Biolabls Inc.), an enzyme that cleaves the bond from the innermost core GlcNAc of an attached oligosaccharide from the asparagine residue (Maley *et al*., 1989). Both pooled male urine and purified protein from pooled male urine were incubated with PNGase F according to manufacturer's protocol and analysed using SDS-PAGE (Figure 5.19). A mobility shift was clearly observed, the cause of which was likely to be due to deglycosylation of the protein. Both resulting bands from SDS-PAGE analysis of the deglycosylated protein, in addition to the urinary protein with attached glycan, were subjected to in-gel digestion. Trypsin and Asp-N were used to give the best chance of producing peptides likely to cover the missing portion of the protein sequence when considering cleavage sites in the existing sequence. All bands digested were identified as the partially sequenced protein (Table 5.2).

Table 5.2| Deglycosylation of pooled male urinary protein from T.vulpecula.
In-gel tryptic and Asp-N digestions of a urinary protein separated by SDS-PAGE, before incubation with PNGase F (C) and after (A and B, separate gel bands), and deglycosylated urine (doublet; A & B) were analysed by LC-MS/MS and identified using PEAKS^{TM}. Peptides identified all bands as the newly sequenced protein.

Sample	-10lgP	Coverage	# Unique peptides
*Control (**C**)*	177.71	34%	10
*Band (**A**)*	238.63	44%	11
*Band (**B**)*	139.53	22%	7

A protocol was then developed to deglycosylate the newly discovered protein under MS-compatible conditions to enable downstream LC-MS analysis of the intact protein, in addition to in-solution digestion with multiple proteases for optimal sequence coverage from LC-MS/MS analysis of the resulting peptides, and subsequent sequence completion. Variations of the manufacturer's protocol for deglycosylation under non-denaturing conditions, which suggested incubation times of between 4-24 hours, and addition of between 2 and 5 µL PNGase F solution, were trialled (Figure 5.20). A band additional to that of the undigested urinary protein was observed after 8 hours with the addition of 5 µL PNGase F, and after 24 hours with both 2 and 5 µL PNGase F, that indicated increased mobility in the gel. However, density of the coomassie staining indicated this additional band was only a small proportion of the total protein analysed. Incubation with PNGase F

under denaturing conditions resulted in a much greater, and more complete, mobility shift than under non-denaturing conditions. However, a reduced mobility shift, as well as resolution into two distinct bands, was observed when pooled male urine was analysed by SDS-PAGE after incubation in the denaturing buffer at 37 °C for 1 hour but without addition of PNGaseF (Figure 5.20, *Denaturing conditions*, 0 units PNGase F). This suggests that the denaturing buffer is responsible for a decreasing shift in mobility when analysing the undigested protein. However, it still indicates that under denaturing conditions, a complete reaction is observed following incubation with PNGase F, whereas under non-denaturing conditions, undigested protein remains even after 24 hours and 2500 units PNGase F.

A protocol incorporating a denaturation step with mass spectrometry-friendly reagents was therefore developed as a more efficient approach to deglycosylation of this particular protein.

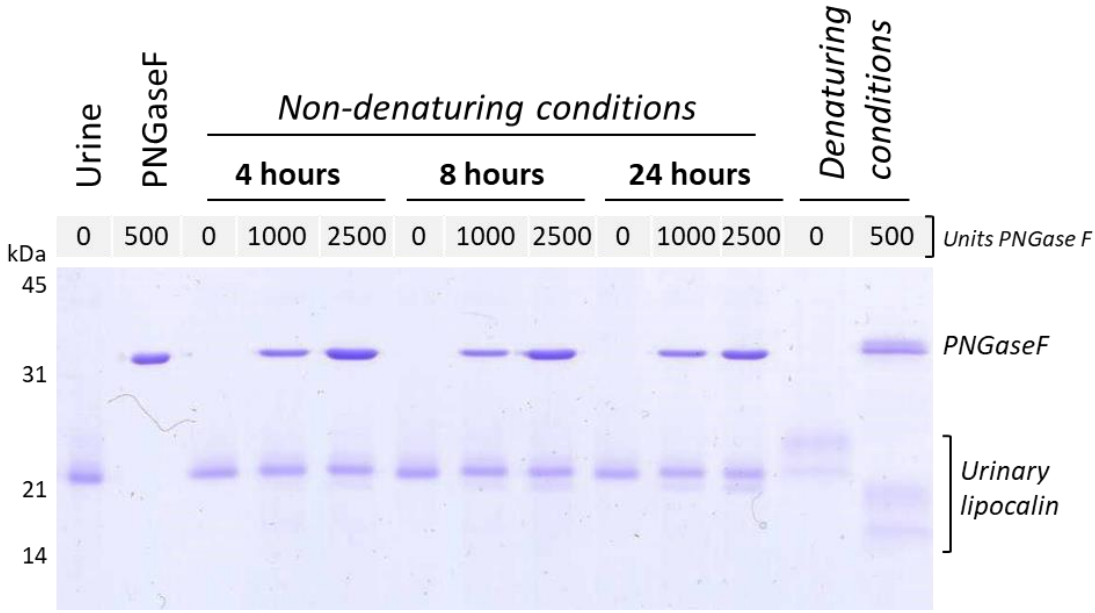


Figure 5.20 | Development of a deglycosylation protocol for downstream mass spectrometry analysis.
Urine was incubated with either 1000 units (2 µL), 2500 units (5 µL) or no PNGase F (-) for 4 hours, 8 hours or 24 hours and analysed by SDS-PAGE. Results were compared with deglycosylation under denaturing conditions with (+) and without (-) 500 units of PNGase F. Brightness and contrast of image was adjusted for clarity.

5.4.5.1 Deglycosylation for downstream LC-MS/MS analysis and completion of sequencing

Different approaches were taken for incorporating a deglycosylation step prior to downstream sequencing by LC-MS/MS and intact mass analysis. For digestion, proteins were previously denatured during the digestion protocol using Rapigest™ SF Surfactant, so incubation with PNGase F was incoporated into the digestion protocol. To ensure the retained activity of PNGase F in the presence of 0.05% Rapigest™ SF Surfactant, pooled male urine was incubated with 500 units PNGase F in the presence of Rapigest™ SF Surfactant or both Rapigest™ SF Surfactant and dithiothreitol, and analysed by SDS-PAGE (Figure 5.21). Under both conditions a mobility shift was observed after incubation. The deglycosylation step was incorporated into the standard in-solution digestion protocol as follows.

Protein (10 µg) was diluted in 50 µL 25 mM NH_4HCO_3 was incubated with RapiGest™ SF Surfactant (0.05% w/v final concentration, Waters, Manchester, UK) at 80 °C for 10 min. The samples were then reduced with dithiothreitol (3 mM final concentration) at 60 °C for 10 min followed by alkylation with iodoacetamide (9 mM final concentration) in the dark at room temperature for 30 min. PNGase F (500 units) was added and incubated at 37 °C for 1 hour. The protein solution was then divided into aliquots for incubation with proteases of different specificity and digested according to 2.5.2.

Aliquots (10 µL) of the protein solution were removed after 1) denaturation and reduction, and 2) after deglycosylation, to check deglycosylation by analysis usingSDS-PAGE. A mobility shift was observed indicating deglycosylation and the peptides resulting from digestion of the deglycosylated protein exposed the N-terminal peptides previously missing from the sequence (see section 5.4.5.3).

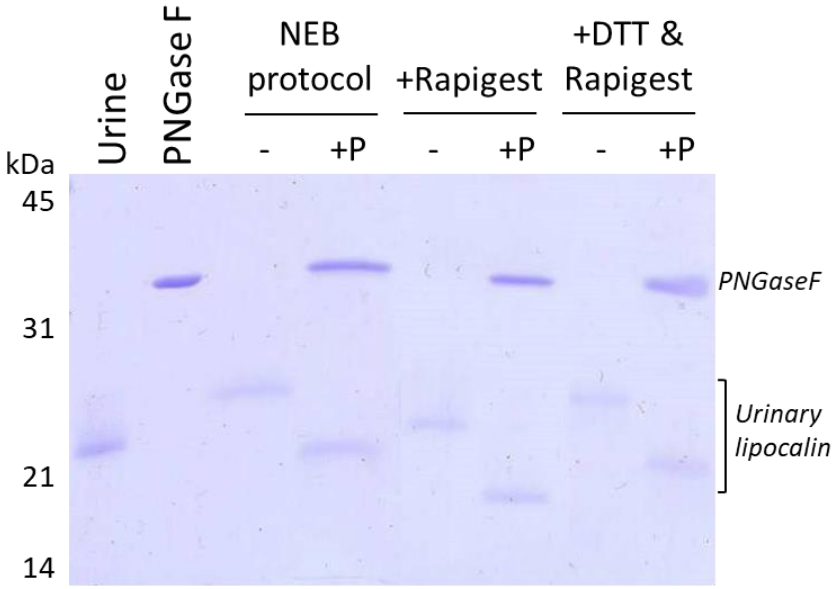


Figure 5.21 | Development of a deglycosylation protocol for downstream digestion and sequencing.
Urine was incubated with either Rapigest™ SF Surfactant (240 µM), dithiothreitol [DTT] (0.5 mM) or both Rapigest™ SF Surfactant (240 µM) and dithiothreitol [DTT] (0.5 mM), in each case incubated with (+P) and without (-) 500 units PNGase F, and analysed by SDS-PAGE. Results were compared with deglycosylation following the manufacturer's protocol (New England Biolabs, NEB) with (+P) and without (-) 500 units of PNGase F (NEB protocol). Brightness and contrast of image was adjusted for clarity.

5.4.5.2 Deglycosylation for downstream LC-MS intact mass analysis

Obtaining an accurate mass for the deglycosylated protein provided an opportunity to confirm the sequencing *de novo*. A protocol was therefore developed to deglycosylate the protein with PNGase F whilst maintaining the intact protein within a detergent-free, MS-friendly buffer. RapiGest™ SF Surfactant (Waters, UK) is suitable for denaturing prior to proteolytic cleavage; during a digestion protocol the detergent is degraded by the addition of acid and the two products removed by centrifugation or reverse-phase liquid chromatography. However, the amount of acid required (final pH < 2) to degrade the detergent also causes most proteins to precipitate. Consequently, an alternative denaturation agent was required. PNGase F is stable in 2.5 M urea at 37°C for 24 h and still possesses 40% activity in 5 M urea (Maley *et al.*, 1989). Urea is a chaotrope commonly used in the denaturation and solubilisation of proteins (Bennion and Daggett, 2003), and can more easily be removed by a buffer exchange step than detergents such as SDS. However, urea can modify proteins: isocyanic acid, a degraded form of urea, induces carbamylation (+ 43 Da) on lysines and arginines and blocks N-termini (Sun *et al.*, 2014). Whilst this is more of an issue for proteolytic digestion efficiency, rather than for analysis of the intact protein

mass, it is still a modification that has the potential to increase complexity of the overall profile. The efficiency of deglycosylation by 500 units of PNGase F in the presence of 5 M urea was subsequently tested and analysed by SDS-PAGE (Figure 5.22). A mobility shift was observed, indicating deglycosylation of the protein.

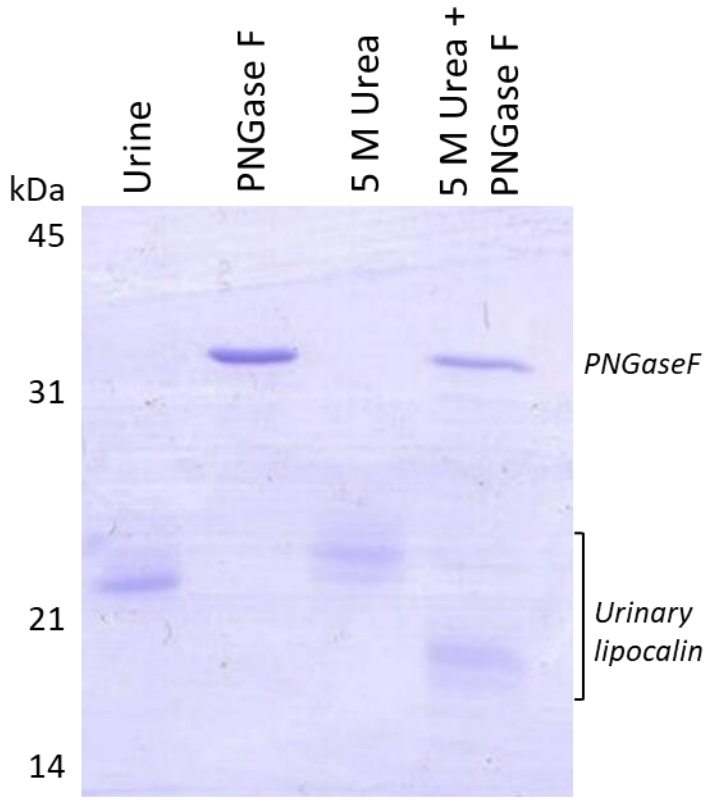


Figure 5.22 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (2).
Pooled male urine from *T.vulpecula* was incubated in 5 M urea for 1 hour at 37 °C, then subsequently incubated with either 500 units PNGase F or a control of water for an additional hour at 37 °C. SDS-PAGE analysis revealed a mobility shift indicative of deglycosylation after incubation with PNGase F. Brightness and contrast of image was adjusted for clarity.

After determining that PNGase F was effective in 5 M urea, the amount of PNGase F required for deglycosylation was assessed. After incubation in 5 M urea, 0, 0.5, 50 or 500 units of PNGase F were added to the reaction mixture and incubated. Subsequent SDS-PAGE analysis revealed 500 units of PNGase F were required to result in a mobility shift indicative of deglycosylation (Figure 5.23).

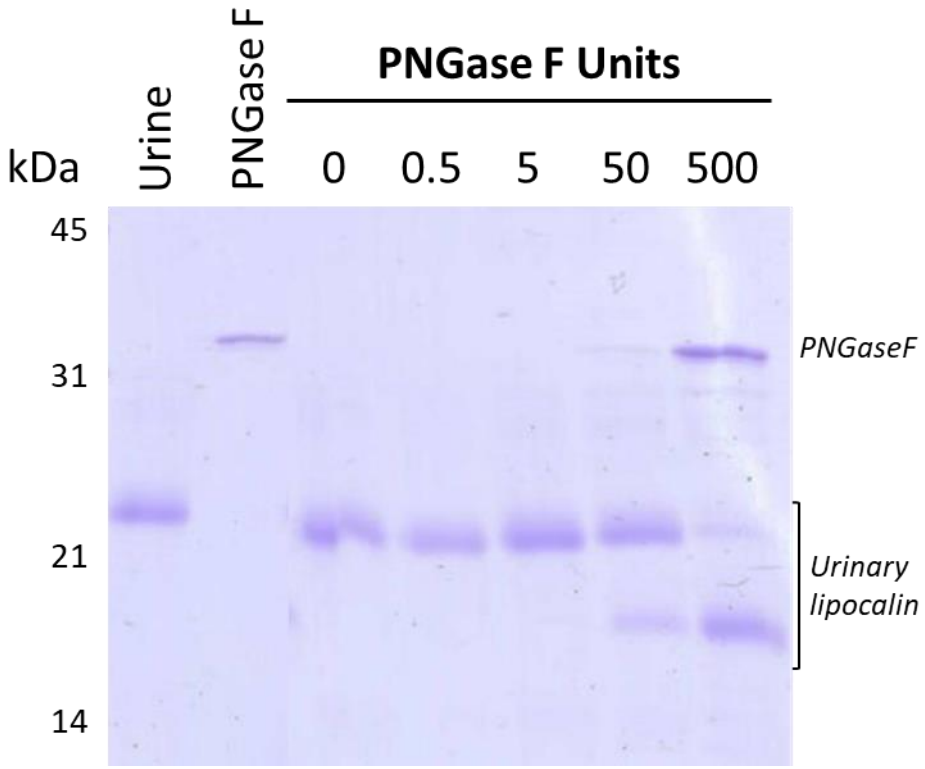


Figure 5.23 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (3).
Pooled male urine from *T.vulpecula* was incubated in 5 M urea for 1 hour at 37 °C, then subsequently incubated with 0, 0.5, 5, 50 or 500 units PNGase F for an additional hour at 37 °C. SDS-PAGE analysis revealed a mobility shift indicative of deglycosylation after incubation with 500 units PNGase F.

For intact mass analysis, 300 pmol pooled male urinary protein was denatured and deglycosylated using the established protocol above. Urea was removed by buffer exchange using Zeba™ Spin desalting columns, and the deglycosylated intact protein was analysed using ESI-MS. A single predominant mass of 18080 ± 1 Da demonstrated a mass decrease of approximately 2200 Da in comparison to the original urine sample (Figure 5.24).

The protein sequence, complete with the newly sequenced N-terminus after deglycosylation (see 'Deglycosylated sequencing') has a predicted average mass of 17962 Da. In the native state, the average mass would be predicted to be 2 Da lighter. The placement of conserved cysteine residues C_{60}, C_{132} and C_{151}, which align perfectly with conserved cysteine residues C_{64}, C_{132} and C_{157} (mature sequence numbering system) in the mouse major urinary proteins (see multiple sequence alignment, supplementary material). In MUPs, C_{64} and C_{157} are oxidised to form a disulfide bond and it is therefore hypothesised that in the newly sequenced protein from *T.vulpecula* that cysteine residues C_{60} and C_{151} do the same. Furthermore, deglycosylation of the protein by PNGase F results in conversion of

the glycosylated asparagine residue to aspartic acid, 1 Da heavier. The predicted deglycosylated mass was therefore 17961 Da. The difference between the observed mass, 18080 Da, and expected mass was 119 ± 1 Da. A search of common modifications of around this mass using UniMod (Creasy and Cottrell, 2004) revealed three potential modifications of this mass; cysteinylation (cysteine residues), pyridylacetyl (lysine or N-terminal) and phenylisocyanate (N-terminal).

A reduction step was consequently added to the deglycosylation protocol to eliminate the possibility of cysteinylation of the free cysteine residue (C_{132}). To investigate the concentration of dithiothreitol (DTT) required for reduction of cysteines in the urinary protein, whilst maintaining activity of PNGase F (which contains three disulfide bonds), three concentrations of DTT were trialled. The denaturing buffer provided by the manufacturer contains 40 mM DTT and the enzyme retains activity, therefore a range of DTT concentrations were selected, the maximum of which was a similar concentration to that of the denaturing buffer (50 mM). Protein from pooled male urine (300 pmol) was incubated in 5 M urea for 1 hour at 37 °C. DTT was added to a final concentration of 0.5 mM, 5 mM or 50 mM for 30 minutes at 37 °C. The protein was deglycosylated with either 0.5, 5, 50 or 500 units PNGase F and again analysed by SDS-PAGE to check the rate of deglycosylation with the addition of DTT (Figure 5.25). Once again, 500 units PNGase F were required to deglycosylate 100 pmol urinary lipocalin.

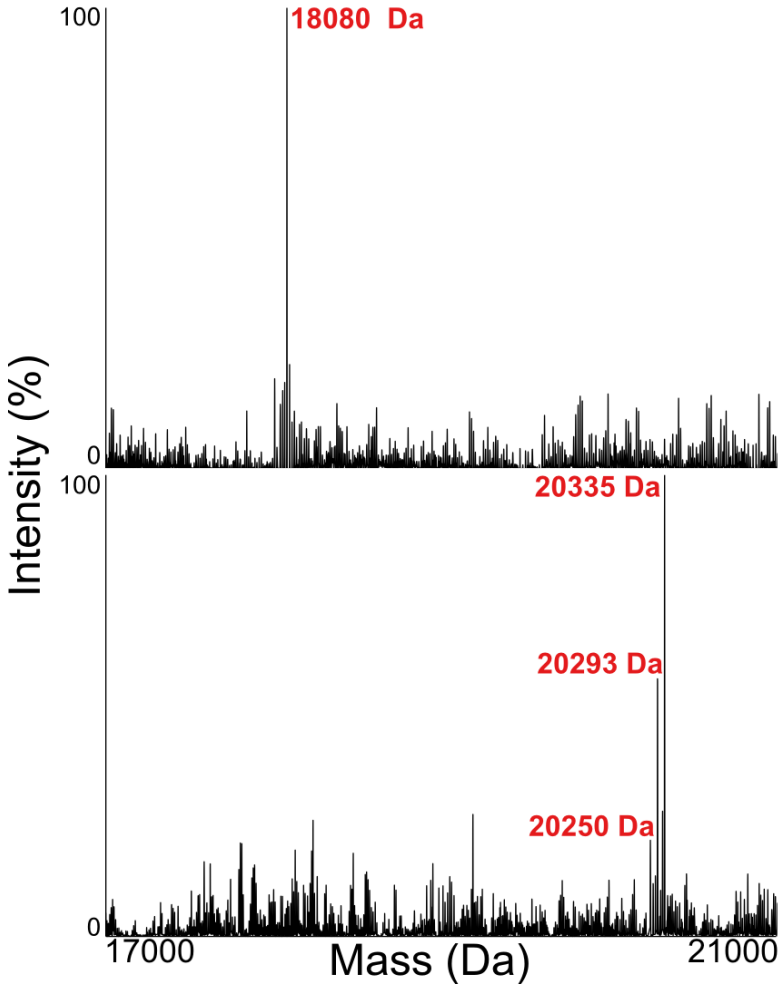


Figure 5.24 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (4).
Pooled male urine from *T.vulpecula* was incubated in 5 M urea for 1 hour at 37 °C, then subsequently 500 units PNGase F for an additional hour at 37 °C. Intact mass analysis revealed a single predominant peak of 18080 Da, approximately 2200 Da smaller than the predominant peaks in the original urine sample.

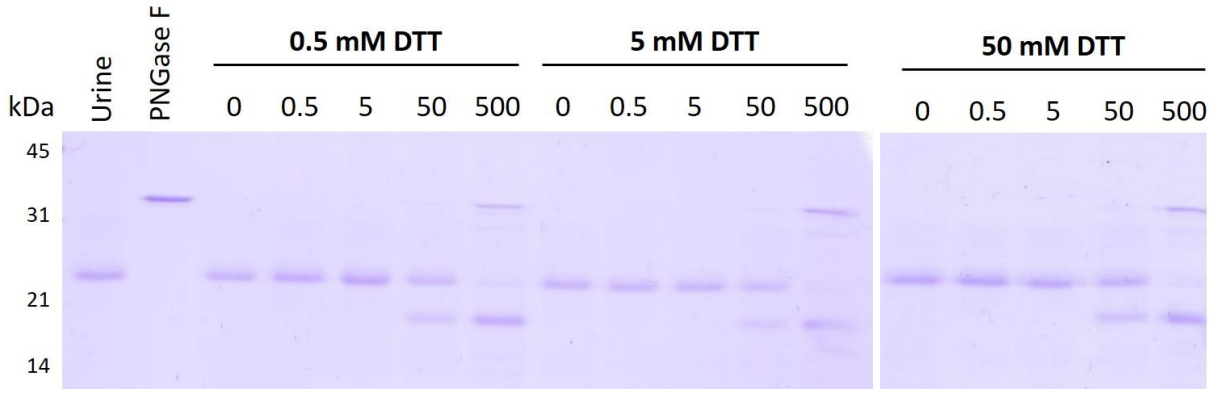


Figure 5.25 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (5).
Pooled male urine from *T.vulpecula* was incubated in 5 M urea for 1 hour at 37 °C, a further 30 minutes after the addition of dithiothreitol to a final concentration of either 0.5, 5 or 50 mM. Finally, 0, 0.5, 5, 50 or 500 units PNGase F were added for an additional hour at 37 °C. SDS-PAGE analysis revealed that under all conditions, 500 units PNGase F were required to see a mobility shift of almost all the urinary lipocalin.

Buffer exchange of the deglycosylated protein under denaturing and reducing conditions was performed using Zeba™ Spin desalting columns and the protein was subsequently subjected to ESI-MS for intact mass analysis (Figure 5.26). Incubation with 0.5 mM DTT did not appear to change the predominant mass peak, however incubation with 5 mM DTT changed the base peak from 18080 ± 1 Da to 17962 ± 1 Da, although many peaks around this mass were also observed. Incubation under stronger reducing conditions of 50 mM DTT resulted in an almost complete mass shift to 17963 ± 1 Da, with only one additional notable peak of 17980 ± 1 Da, 17 Da heavier which is likely to amount to oxidation (16 Da).

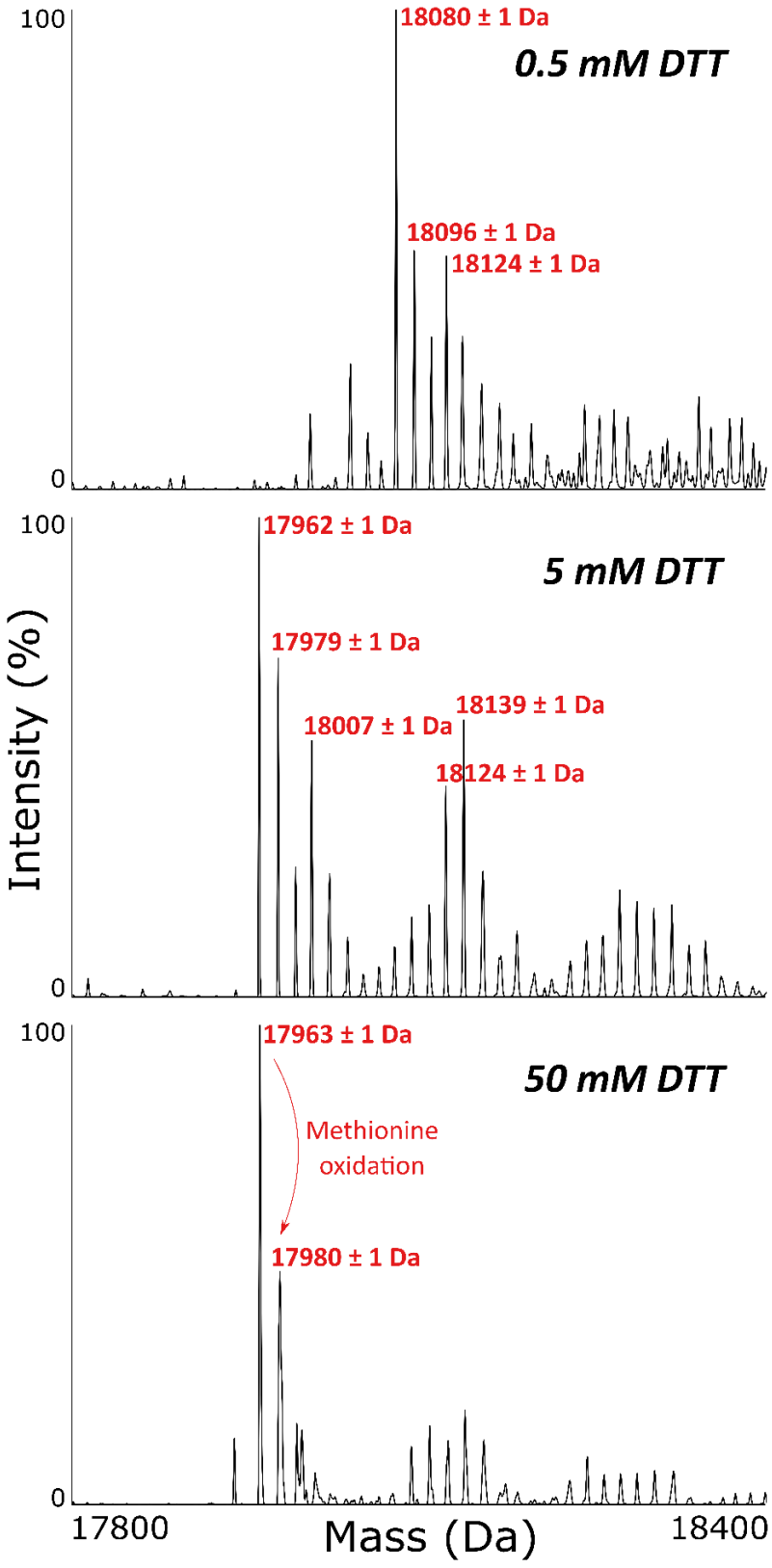


Figure 5.26 | Development of a deglycosylation protocol for downstream intact mass analysis by LC-MS (6).
Pooled male urine from *T.vulpecula* was incubated in 5 M urea for 1 hour at 37 °C, a further 30 minutes after the addition of dithiothreitol to a final concentration of either 0.5, 5 or 50 mM. Finally, 500 units PNGase F were added for an additional hour at 37 °C. ESI-MS was then used to determine the intact mass of the deglycosylated protein.

5.4.5.3 *Completion of the urinary protein sequencing* de novo *after deglycosylation*

Once a deglycosylation step using PNGase F was incorporated into the in-solution digestion protocol, the deglycosylated protein was digested with either trypsin, endopeptidase Glu-C or endopeptidase Lys-C and again analysed by LC-MS/MS. Analysis in PEAKS^{TM} generated candidate sequences that fit within the N-terminal region (Figure 5.28; Figure 5.29). The tryptic peptide JSDNMEDPQMFTGEWFTVAJAS[N>D]VSSK contains the conserved glycosylation motif N-X-[S/T], where the asparagine residue had been converted to an aspartic acid [N>D] by PNGase F (in the PEAKS^{TM} peptide map Figure 5.28 (**C**), this is either indicated by a brown 'd' for deamidation or suggested as a [N >D] mutation). Further, this sequence contains the conserved lipocalin motif G-X-W. It was sequenced from the fragment ion spectra generated from the $[M+3H]^{3+}$ precursor ion of *m/z* 1002.45, in addition to the fragment ion spectra from the modified peptide, oxidised at one or both of the methionine residues present within this peptide (*m/z* 1007.79 and *m/z* 1013.12, respectively). This sequence was confirmed from the Glu-C peptides JSDNMEDPQMFTGE, with one missed cleavage.

The average mass of the protein sequence alone, generated in ExPASy compute pI/Mw tool (Gasteiger *et al.*, 2005) is 17962 Da. When the protein is deglycosylated, PNGase F removes the glycan and the protein mass increases by 1 Da due to the conversion of the asparagine residue to an aspartic acid. The observed mass of the deglycosylated, reduced intact protein at 17963 ± 1 Da (Figure 5.26) therefore supports the sequencing of the novel protein.

The structure of the attached glycan remains to be determined. Glycan components of the glycoprotein, purified by anion exchange, were analysed by Professor Anne Dell and Dr Stuart Haslam (Faculty of Natural Sciences, Department of Life Sciences, Imperial College London), however results were inconclusive. The most abundant masses consistently observed from ESI-MS analysis of the male intact glycoprotein were 20250 ± 1 Da, 20293 ± 1 Da and 20336 ± 1 Da, which after removal of the protein sequence mass of 17162 Da leaves mass deficits of 2288 ± 1 Da, 2331 ± 1 Da and 2374 ± 1 Da. *N*-linked glycans have a tri-mannosyl core attached to the asparagine residue (Gregoire *et al.*, 1996), consisting of two *N*-aceylglucosamine residues, and three mannose residues (Figure 5.27). If it is assumed that the possum proteoglycan has this base structure, this accounts for 910 Da, leaving deficits of 1378 ± 1 Da, 1421 ± 1 Da and 1464 ± 1 Da. Without further analysis it is impossible to determine the remainder of the structure, but as monosaccharide residues

are approximately 200 Da each, it is likely approximately seven monosaccharides are attached to the tri-mannosyl core.

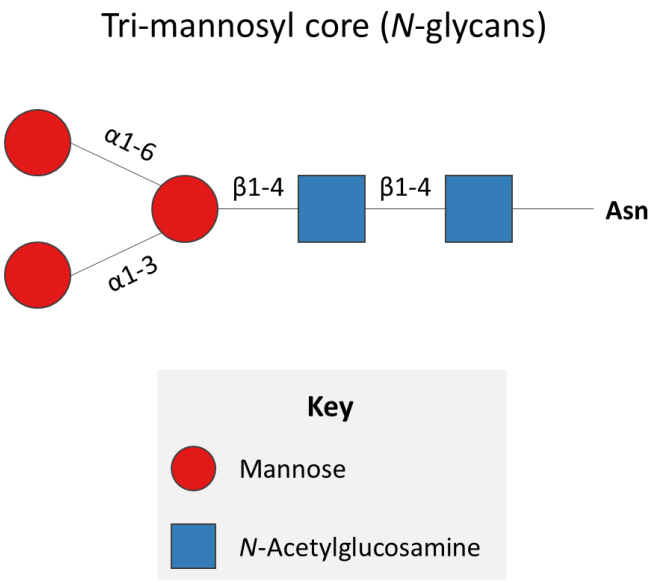


Figure 5.27 | *Structure of the tri-mannosyl core of* N-*glycans.*
Three mannose (red circle) and two *N*-acetylglucosomine residues (blue square) attached to an asparagine (Asn) make up a core tri-mannosyl core structure common to *N*-linked glycans.

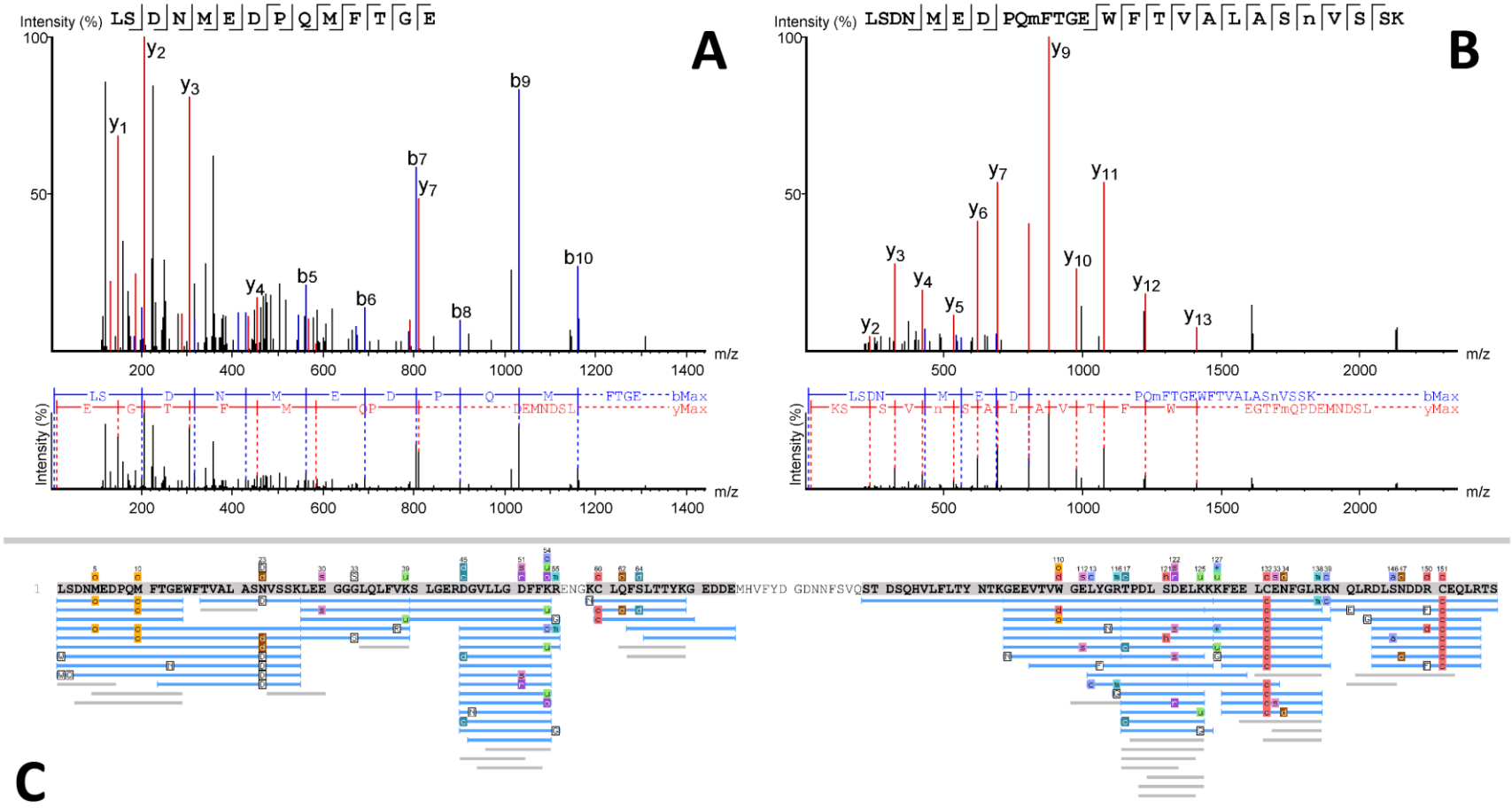


Figure 5.28| ***Deglycosylated sequencing.***

In-solution digestion, using trypsin, glu-C or lys-C, of the pooled male urinary protein was modified to incorporate a deglycosylation step. The resulting peptides were analysed by LC-MS/MS and analysed in PEAKS™. MS2 spectra were manually inspected for two peptides (A, from glu-C digestion & B, from lys-C digestion) that completed sequence coverage of the protein N-terminus.. Peptide data were researched against the novel protein sequence, complete with N-terminal peptides, and peptide mapping showed strong sequence coverage (C).

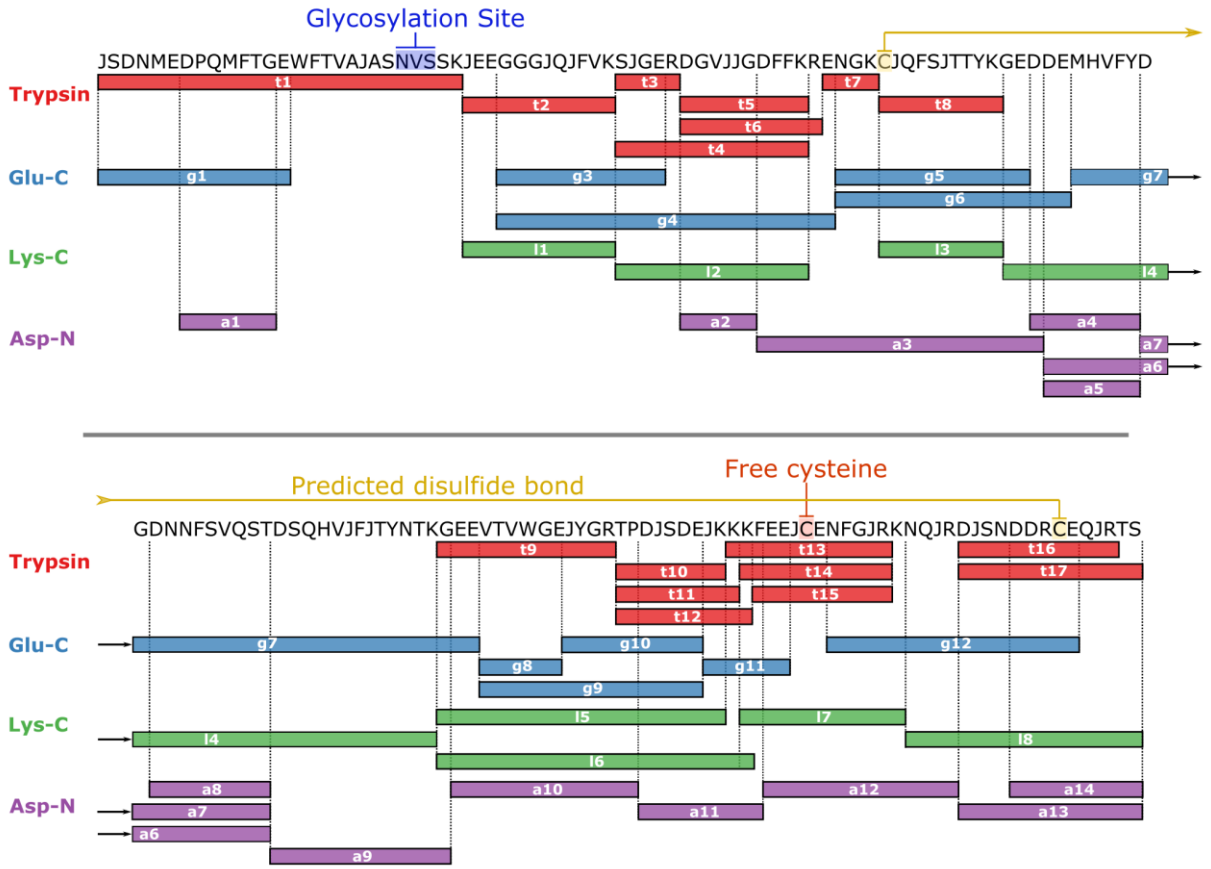


Figure 5.29 | *Protein sequencing*.
The purified protein from pooled male urine was digested with trypsin, Glu-C, Lys-C and Asp-N and sequenced *de novo* with the assistance of PEAKS (Bioinformatics Solutions Inc.). The N-terminus was sequenced from deglycosylated peptides, and sequence coverage of the entire sequence was constructed from peptides derived LC-MS/MS analyses of all proteolytic digestions with and without prior deglycosylation. Overlapping peptides were aligned to form a sequence coverage map that shared homology with other closely-related lipocalin sequences.

5.4.6 Purification, deglycosylation and sequencing of a urinary protein in female brushtail possum urine.

Female brushtail possum urine also contains protein resolving at approximately 25 kDa when analysed by SDS-PAGE, and intact mass analysis revealed a complex protein profile, with multiple mass peaks of approximately 20.2 kDa (Figure 5.30A). To determine if the protein(s) were related to the novel male urinary protein, protein in pooled female urine samples was also subject to purification, deglycosylation and sequencing *de novo*.

Protein was purified by anion exchange chromatography using the same method as for the male protein (see 5.4.3). The manually collected fractions were analysed by SDS-PAGE to locate those containing the protein resolving at approximately 25 kDa. Purified protein was concentrated using Vivaspin® columns and analysed by ESI-MS (Figure 5.30A). Some complexity in the protein profile was lost, but to investigate the protein further, proteolytic cleavage was performed with trypsin, glu-C, lys-C or asp-N and analysed by LC-MS/MS. Data were searched in PEAKS™ against a database comprised of all mammalian sequences in the SwissProt database in addition to the novel sequence. The protein was confidently identified as the same protein present in male urine, which had a protein sequence of mass 17962 Da. A mass deficit between the identified sequence and the intact mass analysis was therefore observed in the same way as for the male protein, and to investigate if this was the result of a glycosylation, the protein from pooled female brushtail possum urine was deglycosylated and subject to further analysis.

Deglycosylation was performed by incubation with PNGase F, with and without the presence of 50 mM DTT with the equivalent protocol than for the male protein, as described in section 5.4.5.2, and analysed by ESI-MS (Figure 5.30; C, deglycosylated; D, deglycosylated and reduced). Whilst the protein profiles of the deglycosylated protein were more complex than that of the male, the smallest mass within the cluster of peaks dominating the spectrum at approximately 18 kDa was 17961 Da, consistent with the mass observed for the male protein. The other peaks observed had masses 17979 Da, 18012 Da and 17995 Da, which differ from the first peak of 17961 Da by 18 Da, 34 Da and 51 Da, respectively. Each increment is therefore 18 Da, 16 Da and 17 Da which, within the constraints of instrument error, could arise from oxidation.

Purified protein from pooled female brushtail possum urine was then deglycosylated prior to in-solution digestion, using the protocol developed in section 5.4.5.1. Proteolytic cleavage of the deglycosylated protein was performed with trypsin, glu-C, lys-C or asp-N

and analysed by LC-MS/MS. Data were searched in PEAKS™ against a database comprised of all mammalian sequences in the SwissProt database in addition to the novel sequence and results were compared to the equivalent analysis of the intact glycoprotein (Figure 5.31). Sequence coverage of the glycosylation motif improved after deglycosylation, and overall sequence coverage confidently identified the same urinary protein sequence as the male brushtail possum urinary protein. The software generated very few putative mutations, which firstly suggests that the female and male urinary proteins are likely identical with no single-point mutations differing between the two. Secondly, it suggests that despite some, albeit low, level of complexity in the deglycosylated protein profile of pooled female urine, it is unlikely to derive from sequence heterogeneity from single-point mutations.

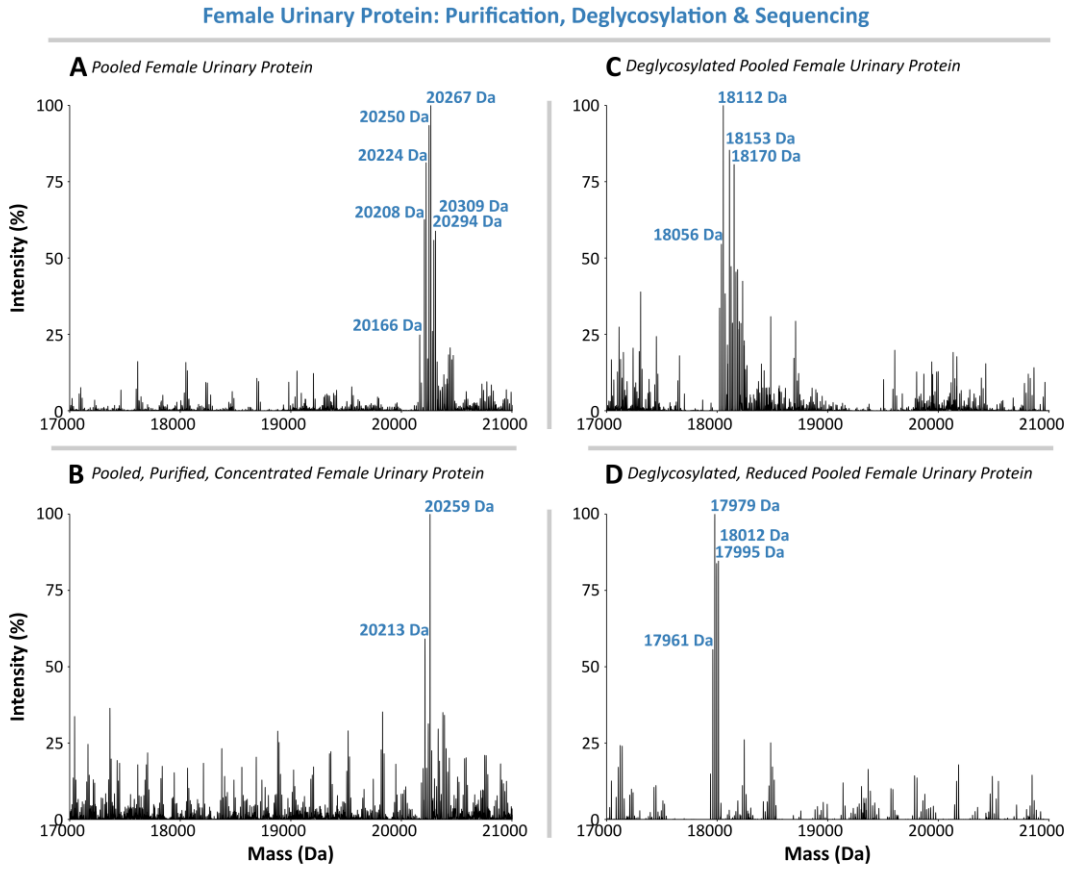


Figure 5.30 | *ESI-MS analysis of female brushtail possum urinary protein after purification and deglycosylation*.
Protein from pooled female urine was analysed by ESI-MS, prior to purification (A) and after purification by anion exchange chromatography and concentration (B).The purified protein was deglycosylated using PNGase F, without (C) and with (D) the presence of 50 mM dithiothreitol, and analysed by ESI-MS.

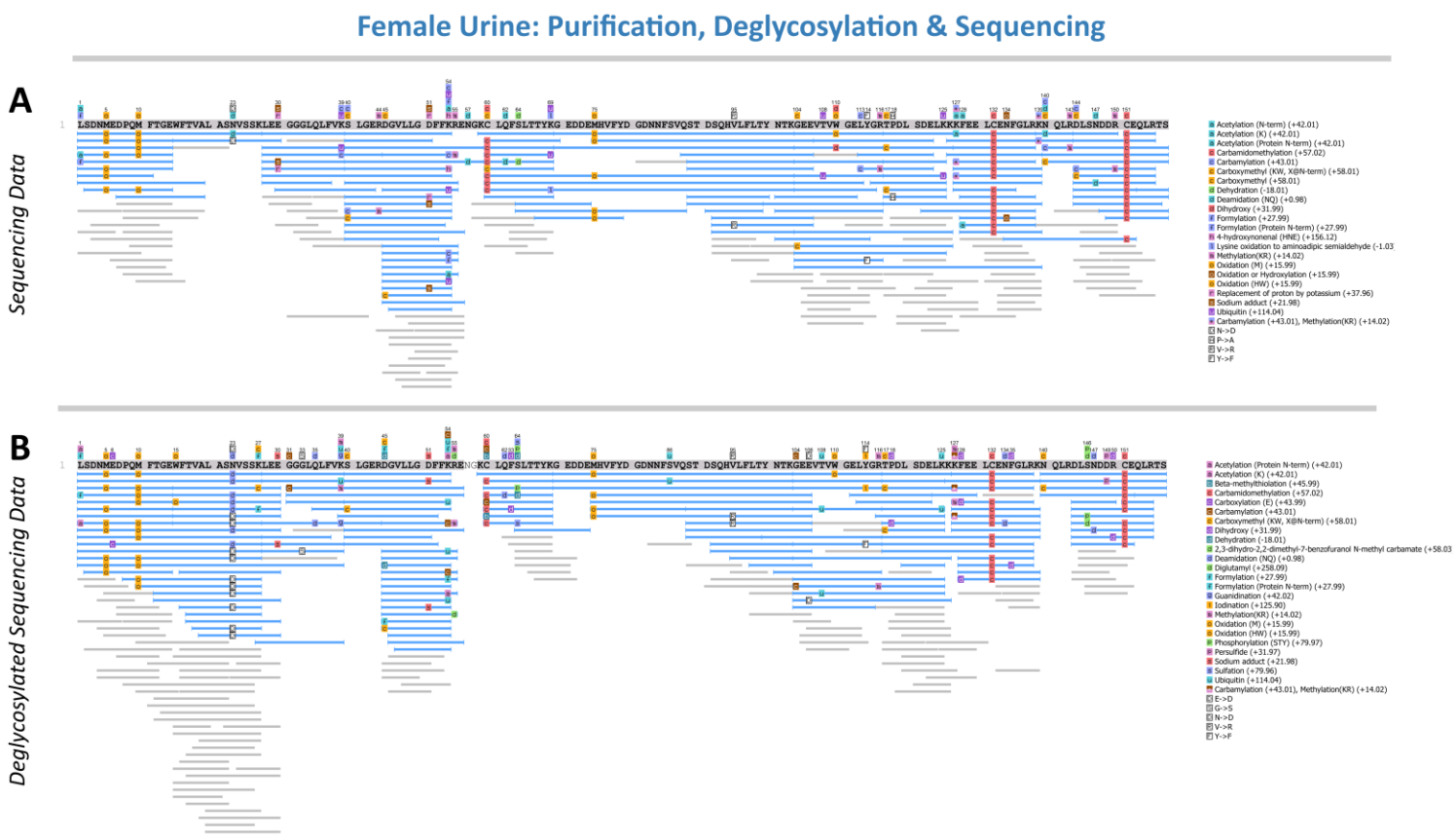


Figure 5.31 | ***Peptide sequence coverage from LC-MS/MS analysis of a purified 18 kDa protein from female brushtail possum urine*.**
Semi-purified urine was digested with four proteases to generate overlapping peptides and the resulting data searched against a database containing the novel male urinary protein to assess coverage. Deglycosylated protein from pooled female urine was also digested with multiple proteases and searched to ensure complete sequence coverage (A-B).

5.4.7 Distinction of leucine & isoleucine residues using isotopic labelling

To complete the sequence, we determined positions of the isobaric residues leucine and isoleucine using metabolic labelling. Isotope-labelled leucine, 5,5,5-[^{2}H] ('heavy') leucine was added to the diet of a male possum over the course of 6 days, to a calculated incorporation rate of 43%, based on dietary consumption. Urine samples were collected prior to incorporation, daily throughout the dietary labelling and for two days once the normal diet was resumed. Samples from each day were digested in-solution with trypsin to monitor heavy leucine incorporation, and the urine sample from day 6 was deglycosylated and digested in-solution with trypsin, endopeptidase Glu-C or endopeptidase Lys-C to provide complete resolution of Leu/Ile-ambiguities. The isobaric residues were determined using two approaches. Firstly, by identifying a +3 Da shift in precursor mass and secondly, by locating the introduction of a +3 Da peak in the **b**- or **y**-ion series of product ion spectra. The incorporation rate was calculated using tryptic peptides containing one instance of leucine. Non-linear optimisation was used to model the difference between the experimental isotope pattern and a theoretical isotopic spectrum generated from a combination of a non-labelled peptide and a 5,5,5-[^{2}H] labelled peptide, the combination of which was dependent on a factor, F. Factor F was consequently used as an estimated value for 5,5,5-d3 incorporation for five peptides (Figure 5.32), which could then be tracked over the course of leucine incorporation.

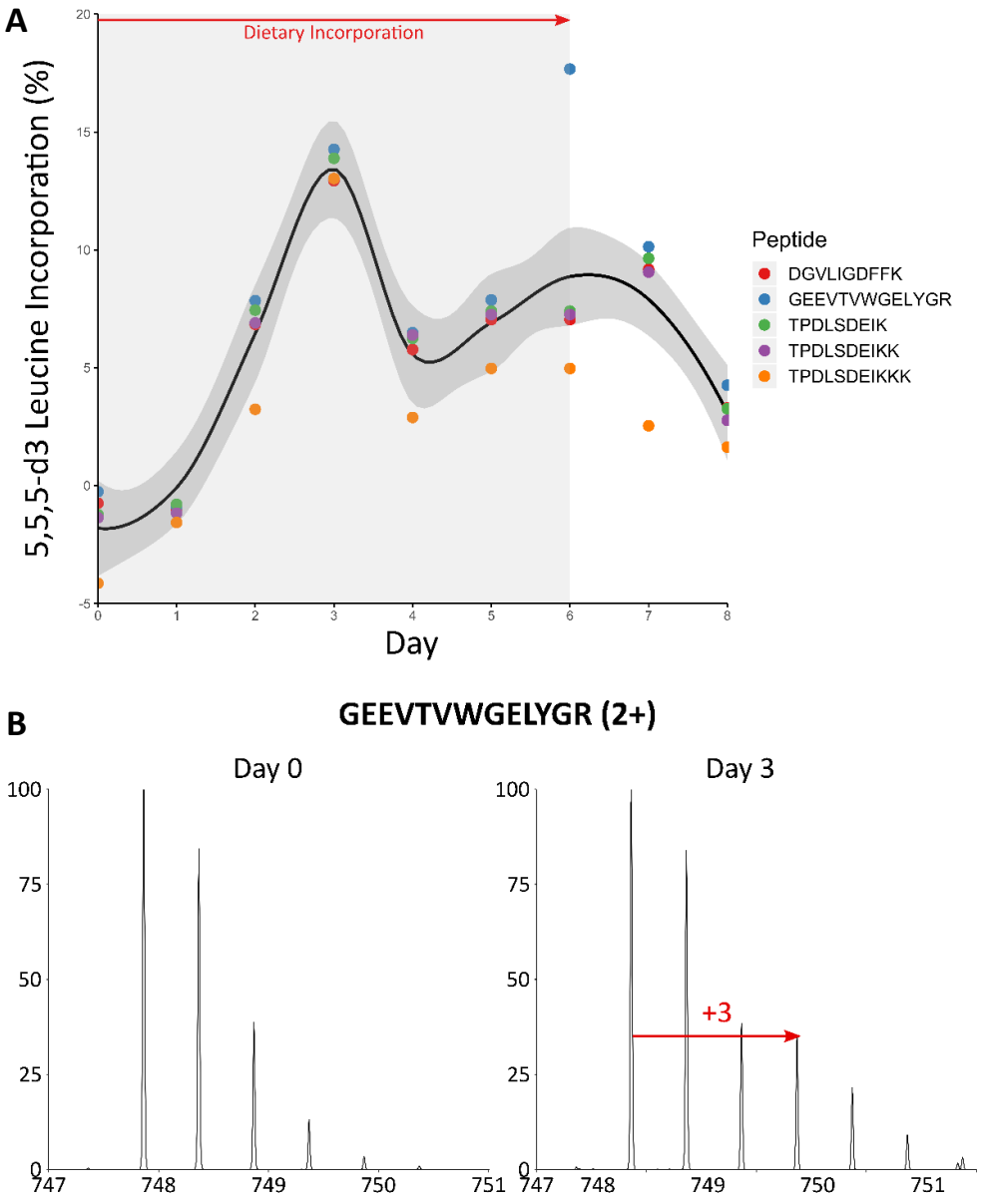
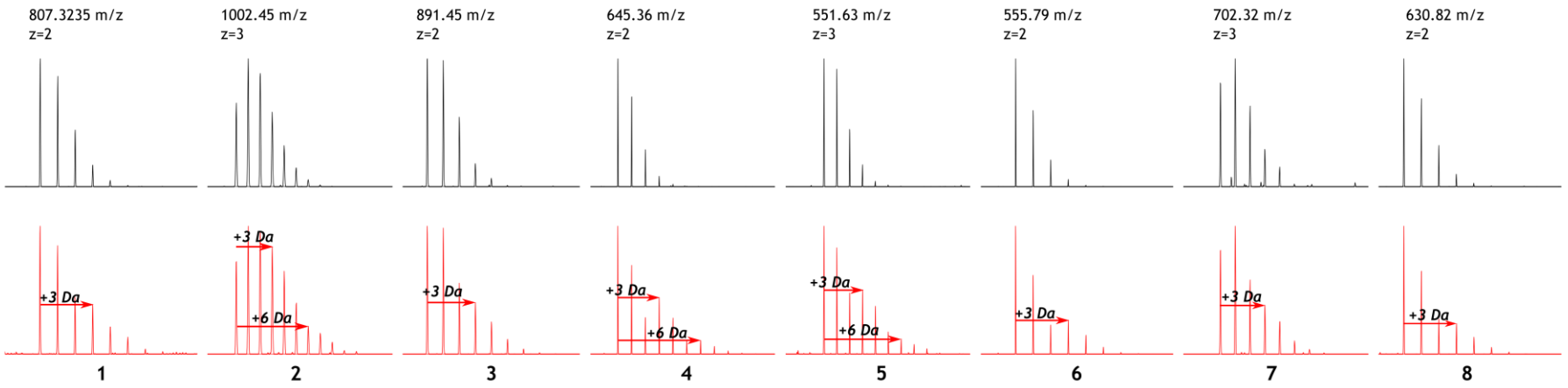


Figure 5.32 | Distinction of leucine and isoleucine using heavy isotope labelling.
Heavy labelled leucine, 5,5,5-d3 L-leucine, was incorporated into the diet of one male possum over 6 days. Five tryptic peptides containing only one heavy leucine residue were used to track incorporation over time (A). Incorporation was highest on day 3 at approximately 15%. An example of incorporation from day 0 to day 3 in the peptide GEEVTVWGELYGR (B) shows a change in the peptide isotope pattern (B).

Heavy-labelled leucine reached a maximum on day three of dietary supplementation at 13 ± 0.5 (mean ± SE) %, however dropped back down to 5 ± 0.5 % on Day 4 and only reached 9 ± 2 % incorporation on day 6, the final day of incorporation. Variation in heavy isotope incorporation rate differs depending on eating habits of the individual relative to sampling time. Despite this, it was still possible to confidently distinguish leucine and isoleucine sites within peptides by using a combination of precursor and fragment ion spectra (Figure 5.33).


807.3235 m/z
z=2
1002.45 m/z
z=3
891.45 m/z
z=2
645.36 m/z
z=2
551.63 m/z
z=3
555.79 m/z
z=2
702.32 m/z
z=3
630.82 m/z
z=2
+3 Da
+6 Da
1
2
3
4
5
6
7
8


LSDNMEDPQMFTGEWFTVALASNVSSKIEEGGGLQLFVKSLGERDGVLIGDFFKRENGKCIQFSLTTYKGEDDEMHVFY
DGDNNFSVQSTDSQHVIFITYNTKGEEVTVWGELYGRTPDLSDEIKKKFEEICENFGIRKNQIRDLSNDDRCEQLRTS
Peptide
Trypsin
Glu-C
Lys-C
Determined by MS2

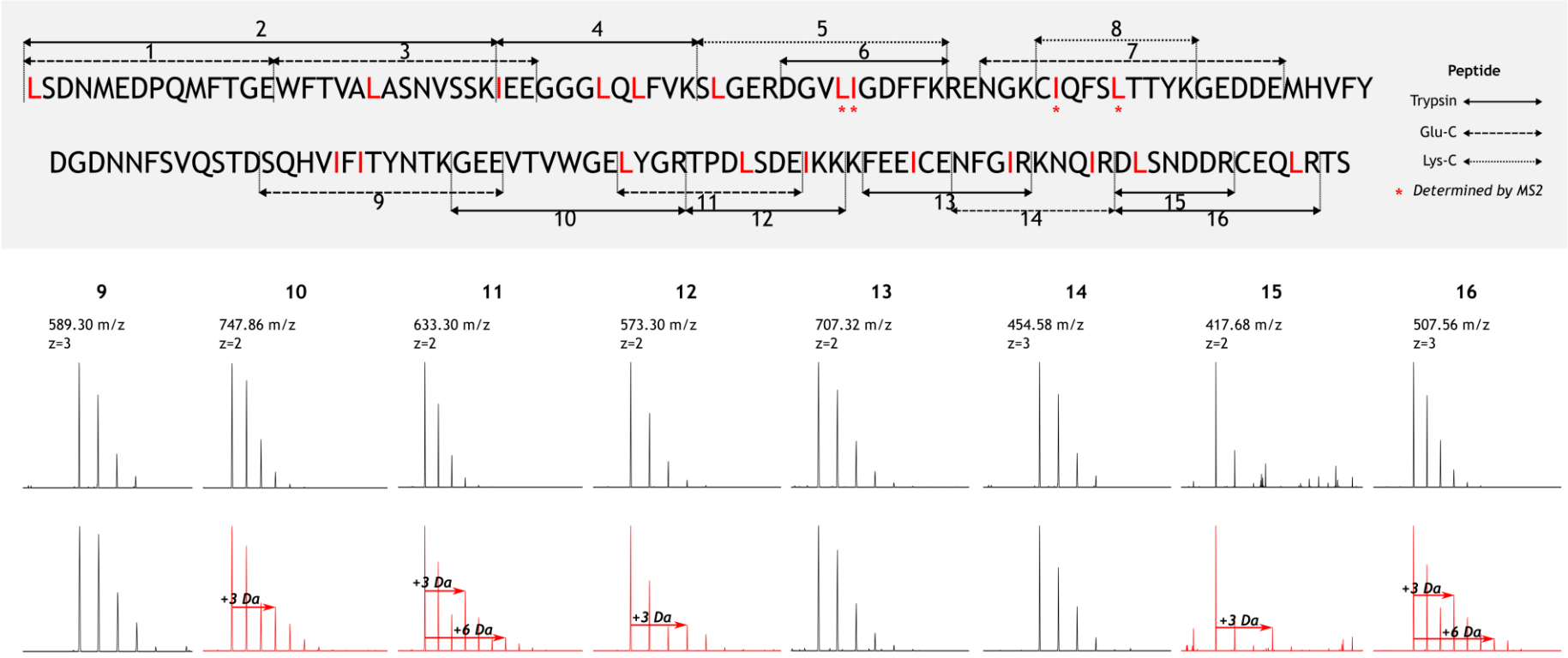


Figure 5.33 | Distinction of leucine and isoleucine using heavy isotope labelling.
Isotope-labelled 5,5,5-d3 leucine was incorporated into the diet of a single male brushtail possum over six days, during which urine was collected daily. Urine samples were subsequently digested in-solution with trypsin, Glu-C or Lys-C (solid, dashed and dotted lines, respectively) to obtain sequence coverage of all Leu/Ile-containing peptides, and leucine sites were assigned according to a +3 Da increase in precursor ion spectra (A-P). Fragment ion spectra were used to determine leucine and isoleucine residues not distinguishable with by MS1 (*), which are found in Supplementary S5.5).

5.4.8 Sequence analysis of the novel brushtail possum protein

Distinction of the isobaric residues leucine and isoleucine allowed in-depth exploration of sequence homology in *Mammalia*, to suggest potential protein function. Marsupial proteins for sequence comparison were identified by searching the finished protein sequence using BLAST® against all marsupial non-redundant protein sequences (taxID:9263) in the NCBI database or all marsupial sequences in UniProt. Sequences from four marsupials were identified as significant (e-value<0.01), including trichosurin from *T. vulpecula*. Sequences from the grey short-tailed opossum (*Monodelphis domestica*), the Tasmanian devil (*Sarcophilus harisii*) and the koala (*Phascolarctos cinereus*) comprised the remaining 30 marsupial sequences. For comparison, lipocalins previously identified as having a putative chemosignalling function were included, the majority of which are from placental mammals. The major urinary proteins (MUPs) from both the house mouse and the Norway rat were included, in addition to salivary lipocalins, epididymal-specific lipocalins and odorant binding proteins (OBPs). For each protein group, reviewed proteins from SwissProt were selected to include sequences from a range of placental mammal species. UniProt and GenBank accession numbers for the final 114 protein sequences used can be found in Supplementary S5.6. Predicted signal peptides were removed using the SignalP 4.1 server (Petersen *et al.*, 2011) and the remaining sequences aligned using Clustal Omega (Sievers *et al.*, 2011). The resulting alignment was viewed in JalView (multiple sequence alignment, see supplementary material) (Waterhouse *et al.*, 2009) and MEGA 6.06 (Tamura *et al.*, 2013) was used for phylogenetic analyses.

Sequence conservation for each residue was explored using JalView (Waterhouse *et al.*, 2009). Figure 5.34 displays sequence conservation of the novel protein sequence within the multiple sequence alignment. Unsurprisingly, the glycine and tryptophan residues in the conserved lipocalin motif [G-X-W] are among those residues most highly conserved. Only one cysteine residue is highly conserved (C_{60}), although cysteine conservation differs between lipocalin groups. For example, the MUPs have three conserved cysteine residues, two of which form a disulfide bond, however OBPs have two disulfide bonds formed from four conserved cysteine residues, despite both groups maintaining a conserved beta-barrel structure.

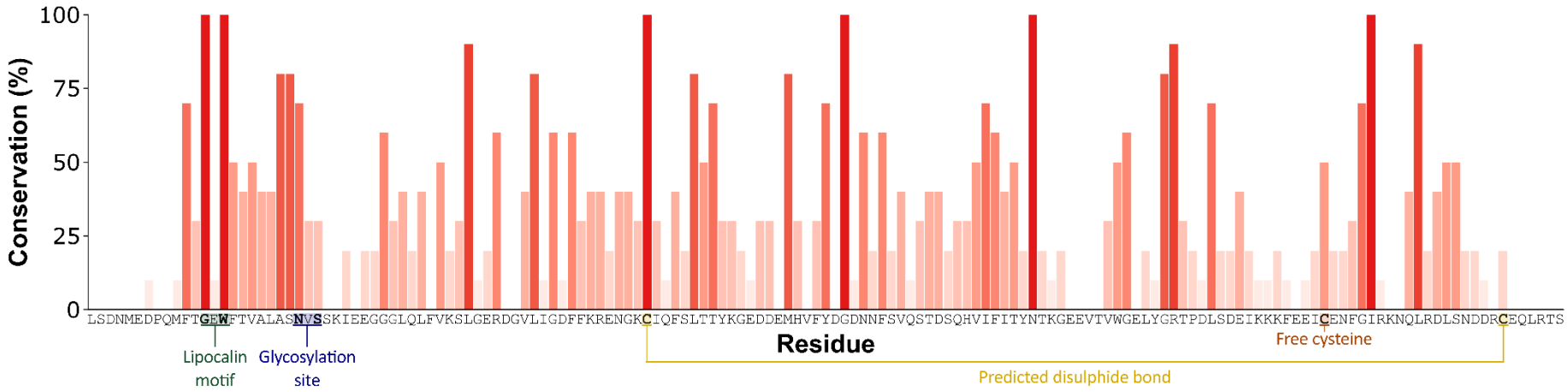


Figure 5.34 | *Residue conservation of the novel protein sequence .*
Conservation (%) of each amino acid position in the novel brushtail possum urinary protein sequence was calculated based as a function of all residues in the multiple sequence alignment. Scores were calculated in JalView 2.10.1 (Waterhouse *et al.*, 2009) from the alignment performed in Clustal Omega (Sievers *et al.*, 2011). Darker bar colour indicates a higher conservation score. The glycine and tryptophan residues of the conserved lipocalin motif [G-X-W] are indicated, as are the cysteine residues.

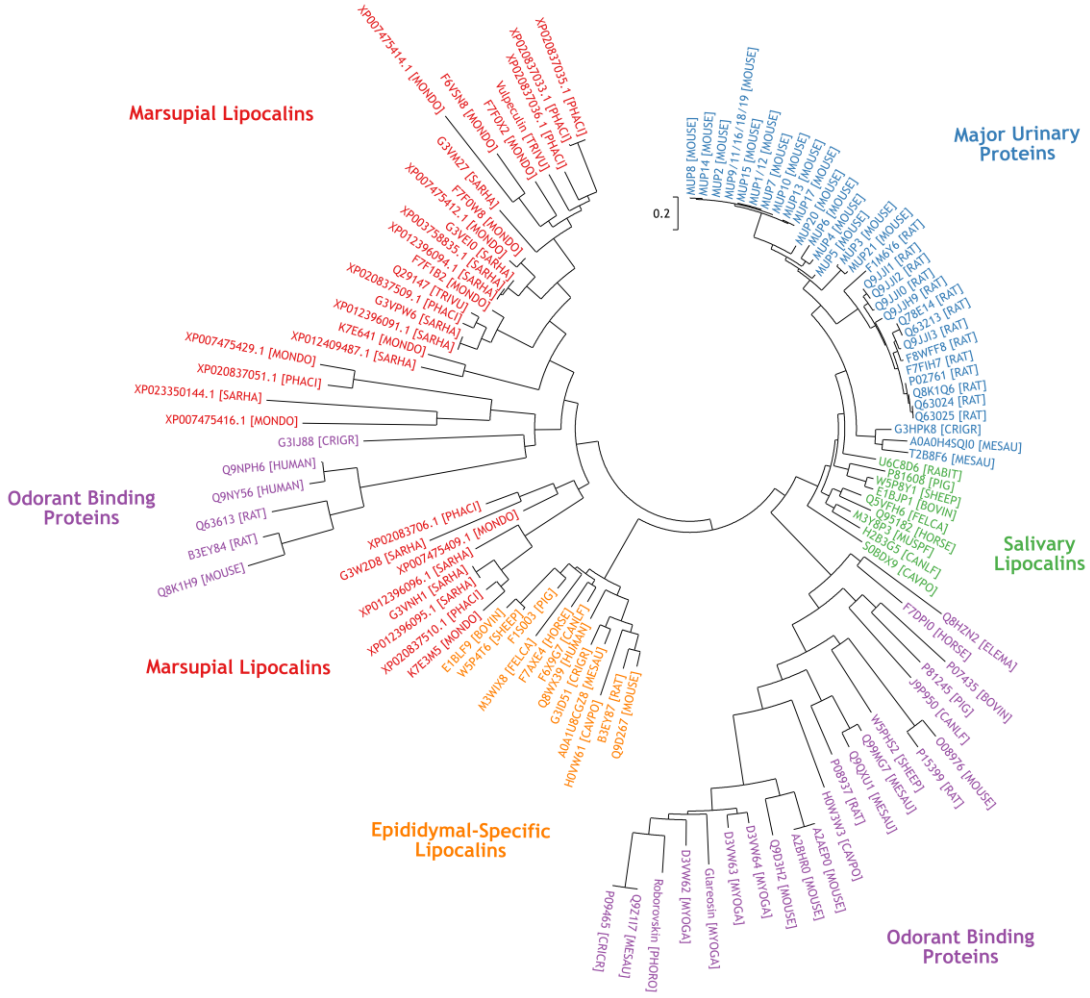


Figure 5.35 | *Phylogenetic analysis*.
The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman model (Whelan and Goldman, 2001). The tree with the highest log likelihood (-16237.8530) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 6.8547)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 114 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 131 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura *et al.*, 2013). The novel protein is indicated with an arrow.

The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan and Goldman model. Bootstrapping analysis using 500 replicates was carried out and the tree with the highest log likelihood (-16237.8530) is shown. Branches corresponding to partitions reproduced in less than 50% boostrap replicates were collapsed. All positions containing site coverage of less than 95% coverage were eliminated

and the resulting 131 positions were analysed in the final dataset. The resulting phylogenetic tree has a number of notable features (Figure 5.35).

The sequences from placental mammals are arranged within their established lipocalin classes. There is clear separation between MUP sequences from the Norway rat and the house mouse, the salivary lipocalin sequences and the epididymal-specific lipocalins. One major clade containing only odorant binding proteins is also observed, although six OBP sequences are located elsewhere in the phylogenetic tree. The marsupial lipocalins (displayed in red), however, group as two distinct clades. One smaller group containing eight sequences are displayed as more evolutionarily distant than the remaining 24. Within these 24 remaining marsupial sequences, four are grouped with the subset OBP sequence outliers. However, the long branch lengths within this clade suggest that this group of sequences is evolutionarily distant within the group. The remaining 20 marsupial sequences are situated in a clade with close branch lengths, and include the novel protein sequence (indicated), in addition to the trichosurin-like protein from *M.domestica* (accession F7F0X2) which was used as an initial sequencing platform, and trichosurin (accession Q29147).

The division between marsupial lipocalins and those from placental mammals suggests that marsupial lineage diverged prior to the separation into the distinct lipocalin classes we see. Consequently, marsupial lipocalins could contain a distinct class of lipocalin, or distinct classes, within those analysed. The naming of the marsupial lipocalins, the majority of which derive from unannotated genomes, makes this distinction unclear. Many are denoted as 'trichosurin-like', 'major urinary protein-like' or indicate homology to the other remaining placental mammal lipocalin classes examined. However, as these marsupial lipocalins are clearly distinct, it would seem more appropriate to re-name these. As distinction of further lipocalin classes within marsupial species is unclear, it was decided not to implement trichosurin in the name of the novel protein. Instead, the naming of the *T.vulpecula* protein followed the same logic as glareosin, from *M.glareolus*, and will be henceforth referred to as vulpeculin.

5.4.9 Structural Homology Modelling of vulpeculin

Whilst the primary structure of lipocalins can vary significantly, the tertiary structure is highly conserved, forming an eight-stranded anti-parallel beta-barrell surrounding an internal calyx that often has ligand-binding properties. To investigate if vulpeculin shares

the same conserved structure, the sequence was subject to structural homology modelling to generate a 3D structure.

All RCSB Protein Data Bank (PDB) structures were searched against the newly sequenced protein using BLAST® (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The six top-scoring sequences (Table 5.3) consisted of MUPs (2A2G, *Rattus rattus*; 1MUP, *Mus musculus*), an OBP (3ZQ3, *Rattus rattus*), salivary lipocalins (5X7Y, *Canis familiaris*; 1GM6, *Sus scrofa*) and trichosurin (2R73, *T.vulpecula*). The sequences were aligned with vulpeculin using Clustal Omega (Sievers *et al.*, 2011) and the corresponding structures used as templates. The alignments were manually adjusted to fit the structural sequence information given in the .pdb file, and 10 models were generated based on each template using Modeller 9.16 (Šali and Blundell, 1993). Model quality was assessed using Ramachandran assessment in MolProbity (Chen *et al.*, 2010), a means of assessing φ and Ψ angles of the protein backbone, and QMEAN score (Benkert, Silvio C. E. Tosatto and Schomburg, 2008), a composite scoring function that assesses the quality estimates of both local residues and the whole structure.

Table 5.3 | *Structural Homology Modelling*.
PDB structures for use as templates for homology modelling of vulpeculin. The 6 top-scoring results from a BLAST® search are listed with scores and pdb accessions.

PDB ACCESSION	DESCRIPTION	MAX SCORE	TOTAL SCORE	QUERY COVER	E VALUE	IDENT
2A2G_A	α-2-u globulin, *Rattus rattus*	108	108	94%	4e-30	40%
3ZQ3_A	Odorant binding protein 3, *Rattus rattus*	104	104	93%	1e-28	40%
5X7Y_A	Lipocalin Allergen Can f 6, *Canis familiaris*	98.2	98.2	96%	4e-26	38%
2R73_A	Trichosurin, *Trichosurus vulpecula*	97.1	97.1	96%	7e-26	39%
1GM6_A	Salivary Lipocalin, *Sus scrofa*	95.5	95.5	96%	3e-25	36%
1MUP_A	Major urinary protein, *Mus musculus*	93.6	93.6	93%	2e-24	36%

The template with the highest average QMEAN4 score was 2R73 (trichosurin) at -1.557 ± 0.0985 (mean ± standard error of the mean). This template also produced the highest scoring model (QMEAN4 score 1.01) which had only 1 outlier residue identified by Ramachandran analysis in MolProbity (Chen *et al.*, 2010), Val_{86}. The only other template to

generate models which had a maximum of one outlier residue was 1MUP (*Mus musculus*). However, the average QMEAN score was -2.235 ± 0.0963 (mean ± standard error of the mean), and consequently, the best fitting model generated from the trichosurin template was chosen. This model of vulpeculin was viewed and annotated in PyMOL (PyMOL, Version 2.1.0).

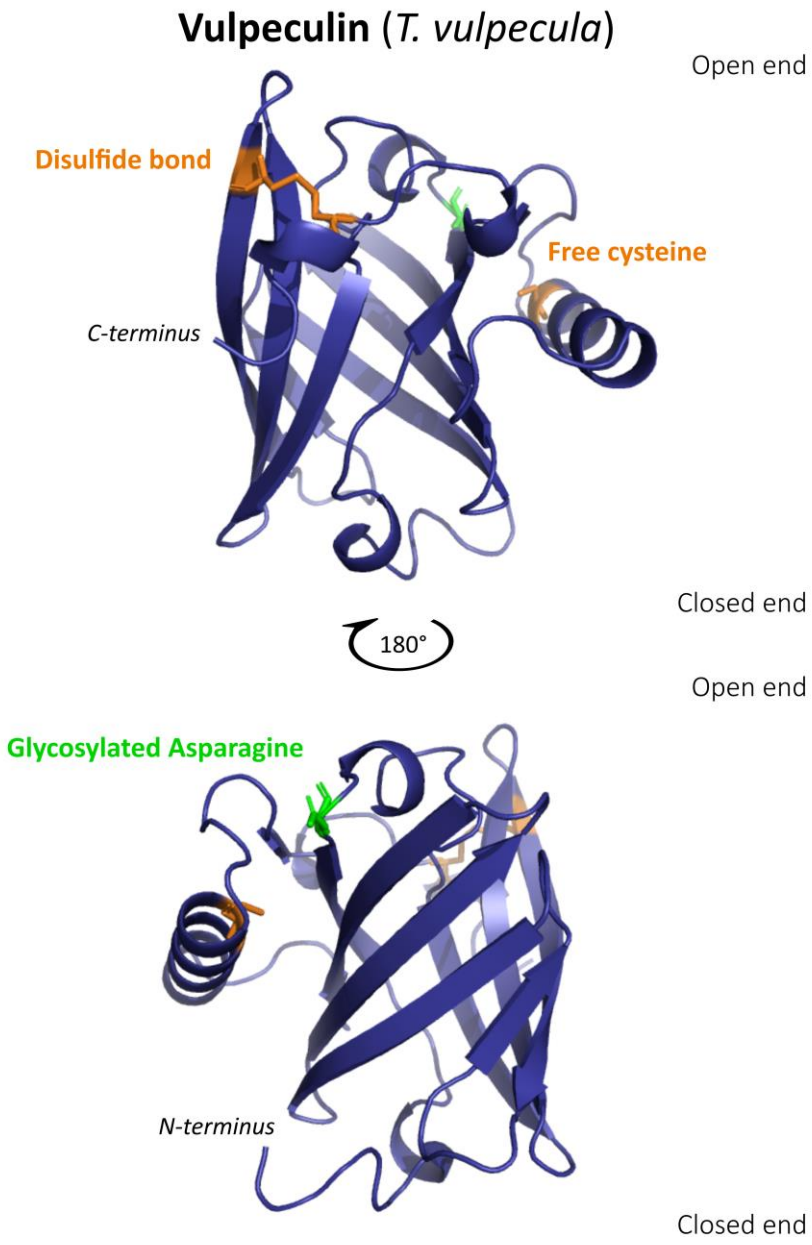


Figure 5.36 | *Structural homology modelling*.
The completed vulpeculin was modelled using homologous templates. Homologous structures were identified in the RCSB Protein Data Bank using BLAST®. Sequences were aligned in Clustal Omega (Sievers *et al.*, 2011) and models were generated in Modeller 9.16 (Šali and Blundell, 1993). Ten models were produced per template, and predicted structures were validated using the QMEAN server (Benkert, Silvio C E Tosatto and Schomburg, 2008) and MolProbity (Chen *et al.*, 2010). The most confident model was built based on the structure of trichosurin (*T.vulpecula*), experimentally determined by X-ray crystallography (Watson *et al.*, 2007).

A conserved beta-barrel structure is observed, consistent with its classification as a lipocalin (Figure 5.36). The glycosylation site (green) is situated at the 'open' end according to lipocalin structure terminology (Flower, 1996). All residues that make up the alpha-helices and beta-sheets were high scoring in the model, indicating these regions are strongly conserved, and as expected, the loops were less confidently modelled. Unfortunately, the attached glycan was not able to be modelled as its structure is not yet determined, but due to its location at the 'open' end of the lipocalin, it may play a role in binding specificity.

5.4.10 Proteome analysis

The predominant protein in *T.vulpecula* urine has been fully sequenced and examined. However, information about other urinary proteins may supplement this information, by predicting tissue origin, or establishing sex differences. Cross-species identification methods of LC-MS/MS data is commonly used in proteomics approaches without the support of an annotated genome. Nonetheless, cross-species matching is restricted by the evolutionary distance of those reference genomes that are available. For example, vulpeculin, which is a very abundant protein in brushtail possum urine, was not identified by cross-species matching. Whilst one factor could be the evolutionary pressure that this particular protein is under, predictably high if it plays a role in chemosignalling, another factor could be the lack of a suitable reference genome; marsupial genomes are not generally well-characterised no fully annotated genome sequence is available for the clade. Results from two database searches were compared to establish the best database to use for cross-species identification of brushtail possum urinary proteins.

5.4.10.1 Database identification comparison

To establish the best database to use to provide the most confident protein IDs, the original whole urine in-solution digests, used to identify the main protein, from seven male and five female samples, was used to compare database hits for a database comprised of all mammalian sequences in SwissProt, and all *M. domestica* sequences in UniProt. A 1% FDR was applied to each database search using the SPIDER search tool. Protein scores were set at a minimum of -10lgP ≥ 20, PTM A Scores were set at -10lgP ≥ 20 and the minimal ion intensity for mutations was 5%. The number of unique peptides required for identification was ≥ 1. Trypsin and other contaminating proteins (for example keratin, actin) were removed from the search results. Most of the protein matches from the database search of *M. domestica* sequence were annotated as 'Uncharacterised'. Each uncharacterised protein

hit was annotated with the description for the closest matching reviewed protein in SwissProt, determined either by examining similar proteins listed in the UniProt entry of the uncharacterised protein, or if unavailable, manually searching the sequence using BLAST within SwissProt.

The score distributions for proteins and peptides were explored, in addition to the sequence coverage and number of unique peptides identified (Figure 5.37). Under the parameters stated above, 84 proteins were identified in the SwissProt database, compared to 57 in the *M. domestica* proteome. However, of those proteins identified, the *M. domestica* database matched 213 peptides in comparison to 156 from the SwissProt database. The score distributions were both significantly different (Log_{10} transformed data) between SwissProt and *M. domestica*, with the *M. domestica* database search scoring higher for both peptide (t-test: $t_{367} = -6.2$, $P \leq 0.001$; difference = -0.077, 95% C.I = -0.10 to -0.05) and proteins (t-test: $t_{138} = -4.1$, $P \leq 0.001$; difference = -37, 95% C.I = -55 to -19). The protein coverage was also significantly higher when searching the *M. domestica* database (t-test: $t_{138} = -2.4$, $P = 0.017$; difference = -3.3, 95% C.I = -6.1 to -6.1), as was the number of unique peptides (Log_{10} transformed data; Mann-Whitney U test: U = 3144.500: $n_1 = 84$, $n_2 = 56$, $P \leq 0.001$; difference = 0.00, 95% C.I = -0.301 to 0.000). Thus, the *M. domestica* database was used to make identifications in whole urinary protein digests. Whilst the number of proteins identified was lower, score distribution for both proteins and peptides was higher when searching against the *M. domestica* database. Whilst score distributions may have been affected by FDR cut-off values, it still remains that the number of peptides identified using the *M.domestica* database was higher, despite a lower number of proteins, and the number of unique peptides was higher, in addition to a better sequence coverage. As a result, the proteins identified were more confident identifications.

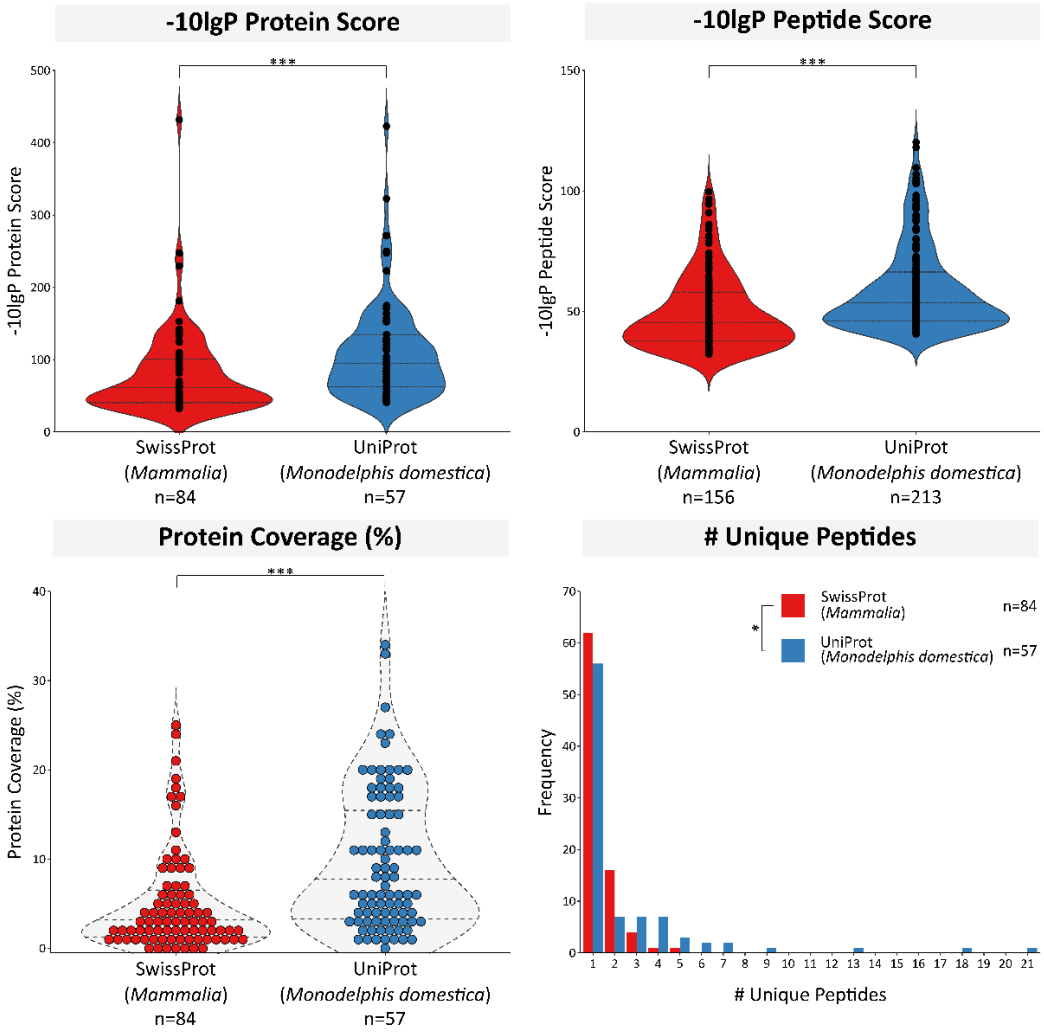


Figure 5.37 | *Comparison of protein identification results when searching peptide data from a global proteomics analysis of brushtail possum urine against a database comprising of all* Mammalia *sequences in SwissProt, or against all* Monodelphis domestica *sequences in UniProt.*

Protein identifications from peptide data of a global proteomics analysis of brushtail possum urine by searching against a database of all mammalian sequences in SwissProt, or all *M. domestica* sequences in UniProt, were compared. Identifications against the *M. domestica* database were significantly higher in terms of protein score (top left) and peptide score (top right). Protein coverage (bottom left) and number of unique peptides (bottom right) were also significantly higher. Violin plots are displayed with median and interquartiles at 25% and 75% data. Statistics were calculated in SPSS. * = $p \leq 0.05$; *** = $p \leq 0.001$.

5.4.10.2 *In-gel digestion of major proteins separated by SDS-PAGE analysis of pooled male and female brushtail possum urine*

Gel bands from SDS-PAGE and Native PAGE analysis of male and female pooled possum urine were isolated and digested for LC-MS/MS analysis and identified using PEAKS™ SPIDER algorithm. In-gel identifications were made by considering both protein identifications and unmatched spectra sequenced *de novo* by the software. The top 10 most abundant unidentified peptides with all individual residues scoring above 50% certainty, sequenced *de novo* by PEAKS™, were manually identified if possible using BLAST searches to assist with identification, excluding trypsin autolysis products and human contamination.

Identifications were dependent on several factors: firstly, the quality of the protein score for matches; secondly, the peptide-based support of these protein identifications, for both peptide spectrum matches and *de novo* tags; thirdly, the abundance of peptides for each protein call was considered; and lastly, that for SDS-PAGE gel bands, the identified protein was acceptable in terms of the expected protein mass. The number and type of each peptide (matched peptide, *de novo* tag or manually assigned peptide identified from unmatched spectra) used to identify each gel band is summarised in Figure 5.38.

The two slowest-resolving bands identified were uromodulin and albumin, commonly seen in urine as discussed in sections 3.4.6 and 4.4.6. Alpha-amylase was also identified, a serum protein commonly found in human urine due to its glomerular filtration (Wu, 2002). One band resolving at approximately 30 kDa was identified as an immunoglobulin, another protein found in urine (Burdon, 1971). None of these proteins suggest that the urine contains any other secretion that could explain the origin of vulpeculin, which was identified from a single band in SDS-PAGE, and from three bands resolving in native PAGE. Under native conditions, one band was considerably more stained than the others in both pooled male urine and pooled female urine. In male urine, the strongest band observed was the fastest-resolving, in females, it was the slowest-resolving of the three. As the protein sequence is the same for both males and females, it is possible that the difference in resolution under native PAGE is driven by the attached glycan.

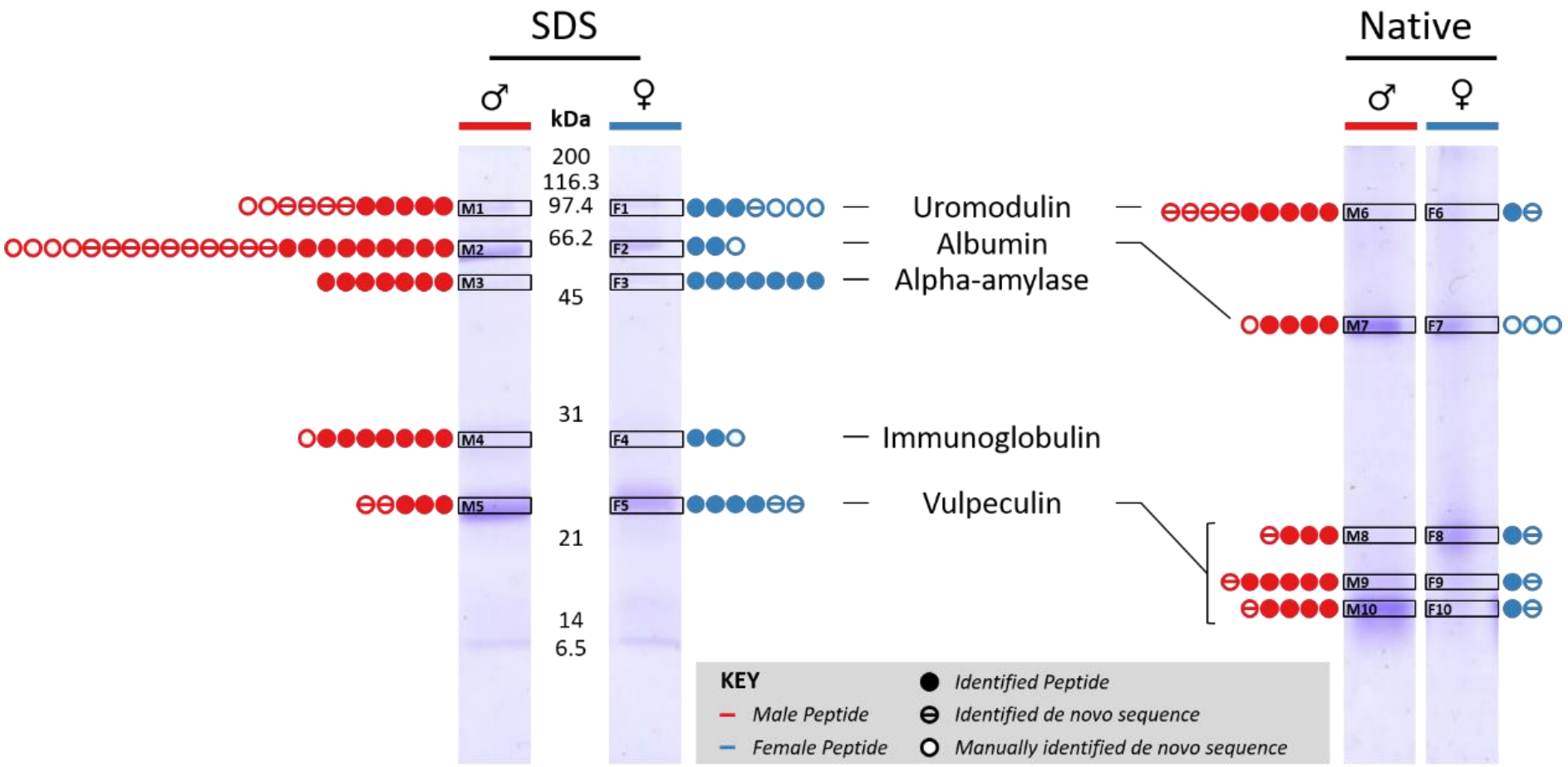


Figure 5.38 | *In-gel identification of brushtail possum urinary proteins*.
Identifications were made by combining sequence data from database searches, in addition to manual inspection of the *de novo* sequences, some of which were identified as *de novo* tags, and some of which were identified using BLAST.

5.4.10.3 Label-free quantification of brushtail possum urinary proteins

Label-free quantification of urinary proteins was used to make comparisons between male (n=7) and female (n=5) brushtail possum urine samples. The protein identifications made from a PEAKS SPIDER search with 1% FDR against the *M.domestica* database were compared. The following parameters were set according to the data; a minimum peptide quality score of 5, and a minimum feature area was set to 1 e^5. A minimum peptide count of 3 was set to calculate abundances, and the number of confident unique peptides was 1. Significance was calculated using the PEAKSQ method although no filter was applied.

After filtering, trypsin and human contamination proteins were removed from the analysis to leave 16 proteins remaining, including vulpeculin. Of these, five had a fold change of 2 or more (with respect to a reference sample automatically selected by the software), above a significance level of 25 (recommended) (Figure 5.39). These were all increased in male samples, and included vulpeculin, which increased 9-fold in males. The other proteins significantly increased in male samples were serum albumin, alpha-fetoprotein, uromodulin and cadherin-1. The presence of these proteins, particularly albumin and uromodulin, is common in urine and no significant increase in either male or female samples should be observed, and suggests that normalising to creatinine may introduce a bias rather than normalising for urine dilution.

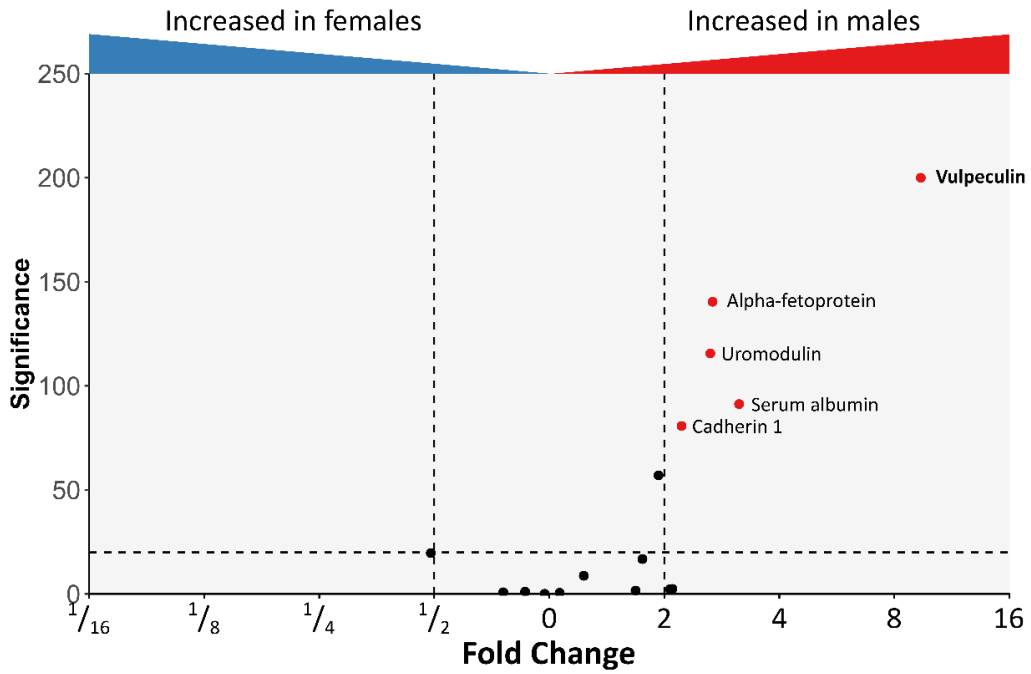


Figure 5.39 | *Label-free quantification*.
Protein abundances were compared using label-free quantification in PEAKS™. Log_2(Fold Change) is compared with protein significance. Proteins with a significance higher than 25 and a fold change of more than two are indicated.

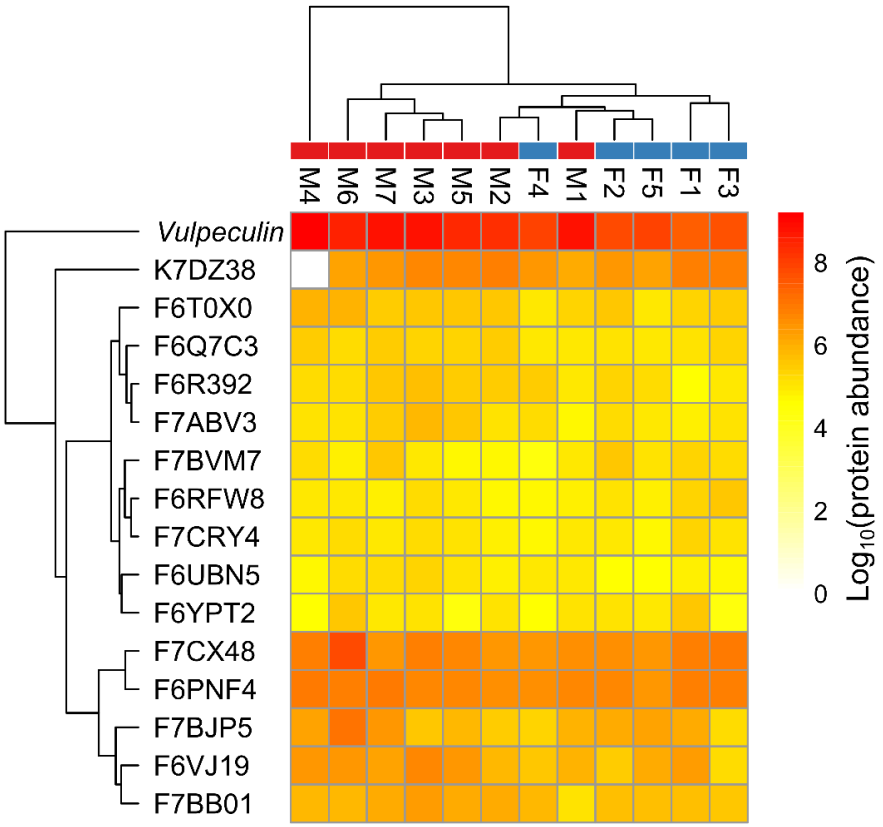


Figure 5.40 | *Label-free quantification*.
Cluster analysis was used to determine similarities between individual male and female samples. Both protein accession and sample were clustered and Log_{10}(Protein Abundance) is displayed as a heatmap.

Correlation analysis was used to investigate if, given the relative quantities for the 16 proteins, samples would be grouped by sex, and if other proteins displayed a similar protein abundance pattern to vulpeculin. Figure 5.40 Log_{10}(normalised abundance) values as a heatmap, where both protein accession and sample are subject to clustering . No other proteins were clustered with vulpeculin. However, four male samples clustered together (M3, M5, M6 & M7). An outlier sample was male 4, which may have been driven by the absence of identification for K7DZ38 (neurocan core protein). Two female samples (F1 & F3) clustered together, however three were clustered with two male samples (F2, F4 & F5). Hierarchical clustering therefore suggests that relative abundance of urinary proteins is not enough to fully differentiate between sex.

Overall, protein identification was poor. Only a small portion of the acquired spectra were identified. Many more spectra were of high enough quality to be sequenced by the software, however the candidate peptides generated were not identified by cross-species matching.

As protein identification was poor, a feature analysis was performed in Progenesis QI to explore if differences in the whole dataset were driven by sex. A principal component analysis (PCA) plot of all peptide data demonstrated that samples from members of the opposite sex was clearly separated by principal component 1 (Figure 5.41). This confirms that the major factor in data variation is sex.

A

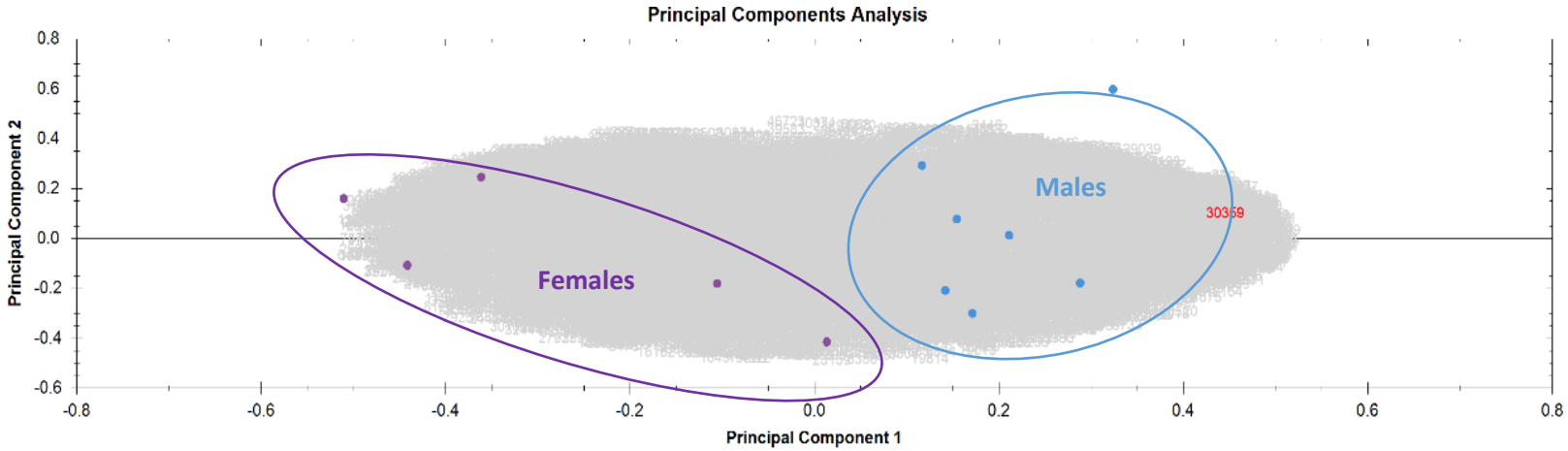


Figure 5.41 | *Label-free quantification*.
Peptide features were analysed in Progenesis QI to explore causes of major differences in the data. Male (blue) and female (purple) samples are separated by principal component 1, even when a q-value of ≤ 0.05 was not applied (A), in addition to when it was (B).

5.5 Discussion

A novel glycosylated urinary protein was identified, sequenced and characterised from the common brushtail possum (*T.vulpecula*). A predominant protein in both male and female urine samples, it was semi-purified, isolated and sequenced. A mass deficit between the sequence and the intact measurement of the protein was concluded to be the result of an N-linked glycan. After removal of the glycan, the remainder of the protein was sequenced, and completed by determining the isobaric residues of leucine and isoleucine using heavy isotope-labelled leucine incorporation to the possum diet. The relationship of the completed sequence to homologous proteins was explored, and the protein defined as a member of the lipocalin family. Structural homology modelling also showed a good fit to the conserved beta-barrel structure of lipocalins. An attempt at identification of other proteins in *T.vulpecula* urine was explored firstly by in-gel identification, and secondly by label-free quantification; however quality identification was hindered by low levels of cross-species matching.

Overall analysis of urine was assessed by SDS-PAGE, protein concentration relative to urine dilution, and intact mass analysis of samples obtained from individual possums. Both SDS-PAGE and overall protein concentration (Bradford assay) suggested a sex difference in the overall protein output and the abundance of a protein resolving at 25 kDa, however this abundance difference was not statistically significant. Whilst the appearance of a sexual dimorphism is not as distinctive as that of the bank vole, *M.glareolus*, it could perhaps be worth investigating further. The female protein concentration could be skewed by one individual, whose protein output was consistently higher than the others, although the animal of origin was taken into account in the mixed effect model analysis. Future analysis could benefit from sampling a greater number of animals however this analysis was restricted by logistical considerations. Another angle that was not considered is the effect of season. Sampling took place year-round and no difference was observed between those samples collected in breeding and non-breeding seasons, but additional measures were not taken in this preliminary investigation to sufficiently investigate the breeding status of animals. The breeding season of the captive brushtail possum is observed twice a year, however given how the breeding season of these animals can fluctuate (Efford 1998; Jolly et al. 1995), further biological controls such as testosterone levels or oestrus status could provide better confirmation.

Intact mass analysis determined that the dominant protein in the majority of samples had an intact mass between 20 kDa and 21 kDa. However, the disparity in masses determined between samples has still not been fully resolved. The quality of spectra acquired differed between samples, despite normalising to overall protein concentration. Spectral quality will be influenced by two main factors; by the proportion of each protein within the sample and the level of contamination. In samples where the investigated protein occupies a higher proportion of the total protein output, the intensity of *m/z* peaks for this protein will be higher. Abundance of vulpeculin as a proportion of total protein output was a primary issue when analysing female samples. Furthermore, in the urine samples from *T.vulpecula*, a pattern of incremental masses was observed which suggested modification of the protein, or proteins. This modification not only masked the level of heterogeneity at the primary protein sequence level, but the increased number of protein forms meant that the total signal was split across the different species and signal intensity was compromised. The origin of the modifications is also unknown; it is unclear if these modifications occur biologically, or during sample handling and freeze/thaw cycles, or due to the buffer used for ESI-MS analysis. It is therefore impossible to know if the repeated 42 Da increments are derived from modification of the glycan, or from different species of the glycan itself. Ideally, these issues would be explored further, however it is enough to say that in consideration of the aims of this project, as a preliminary investigation into the proteomic content of *T.vulpecula* urine, that while the protein component of this molecule appears to be consistent, the attached glycan is an unexplored source of potential heterogeneity. An ideal additional control would also be the intact mass analysis of the deglycosylated protein in individual samples..

The difficulty of developing a protocol for glycan removal for ESI-MS was due to disruptive conditions required for effective and complete removal, within the constraints of MS-friendly conditions. An incubation time of 24 hours with up to 2500 units PNGase F under non-denaturing conditions only showed a small change in mobility on an SDS-PAGE gel. This in itself suggests that the overall structure of the protein-glycan complex may inhibit access of the enzyme to the glycosylated asparagine residue. Longer incubation times could be an alternative, but long-term exposure of intact proteins risks disruption to tertiary protein structure, although lipocalin proteins are known for their robustness. An MS-friendly denaturation method was chosen, however this relied on balancing the ability to denature vulpeculin enough for enzymatic cleavage of the glycan, yet still maintaining activity of PNGase F. Many detergents and organic solvents usually used for denaturation of proteins

are incompatible with either MS analysis or PNGase F activity, so urea was chosen. However, the spectral quality was relatively low, and improvements to the protocol would ideally be made before applying it to individual samples, particularly as the protocol consumes relatively large quantities of PNGase F (for reference, 1 unit is the amount of enzyme required to remove > 95% carbohydrate from 10 µg denatured RNase B in 1 hour at 37 °C. Comparatively, the developed protocol used 500 units for 1 µg protein). Deglycosylation of the semi-purified protein instead of using whole urine was also considered, however with regard to the amount of urine sample used in method development, purified protein would have been costly in terms of both sample and time.

An attempt to sequence the glycan was made, in collaboration with Professor Anne Dell and Dr Stuart Haslam (Faculty of Natural Sciences, Department of Life Sciences, Imperial College London). Despite good glycan identification data (not shown), it was not possible to match the data with the intact mass deficits observed. Semi-purified protein from pooled male and female samples (originally used to sequence the protein) were analysed. It could be that the protein was not purified enough, or that there was further complexity from modifications that made it difficult to confidently match glycan residues. Although many of the intact masses were similar between samples, it was also not possible to confirm if the glycan had the same basic structure in individuals. Ideally, glycan analysis would be performed on completely purified protein from individual samples, however the purification required was too time consuming to perform, with too little yield for effective analysis. At present, only the protein component is confidently sequenced.

Homology analysis of the completed protein sequence provided considerable insight into evolutionary relationships within the lipocalin protein family. Phylogenetic analysis revealed that marsupial lipocalins identified as homologous to the novel protein did not cluster into the well-described lipocalin families of eutherian (placental) mammals. As the described protein was situated within a cluster of marsupial lipocalins, and the functionality of these sequences has not been previously explored, the novel protein was named vulpeculin, rather than likened to a previously defined lipocalin group from eutherian mammals. The phylogenetic analysis performed here aimed to characterise the relationship of the novel vulpeculin sequence to other lipocalin classes known to have roles in chemosignalling. However, further analysis of a more expansive range of lipocalins, including lipocalins identified in marsupial milk (for example, trichosurin), may provide further insight into the divergence of this class of marsupial lipocalins. The closest inferred homology was to other marsupial sequences identified using BLAST®.

Although the genome sequence for *T.vulpecula* is unavailable, investigation into the location of genes encoding proteins from other marsupials reveals a cluster of lipocalin-encoding genes. A species-specific expansion of both vomeronasal-2 receptor genes and MUP-like genes has previously been reported in the grey short-tailed opossum, *M.domestica*, from which parallels are drawn to the corresponding expansions in mice and rats (Chamero *et al.*, 2007). However, it has been noted that again, these sequences were sufficiently divergent from placental mammal MUPs to suggest classification within an alternative lipocalin class (Logan, Marton and Stowers, 2008), and the same grouping of marsupial lipocalins, distinct from other mammalian sequences is also paralleled in β-lactoglobulins (Piotte *et al.*, 1998). The cluster of lipocalin genes on chromosome 1 of the *M. domestica* gnome encodes at least six lipocalin proteins (UniProt accessions: F6VSN8, F7F0X2, F7F1B2, F7F0W8, K7E641, all included in the analysis, and K7E3M5, not identified by BLAST searching so not included). Less published information is available for *S. harrisii* (Tasmanian devil) and *P. cinereus* (koala), although tBLASTn® searching of the vulpeculin protein sequence identified a number of lipocalin-encoding genes located together in *S. harrisii* (encoding for G3VM27, G3VPW6, G3VEI0 (UniProt Accession), XP_012396096.1 (NCBI Accession), and G3VNH1, denoted here as XP_012396095.1) and in *P. cinereus* (encoding for XP_020837033.1, XP_020837035.1, XP_020837036.1, XP_20837509.1 and XP_020837510.1, NCBI Accession). This suggests these proteins may be co-expressed, and it may be the case that some level of polymorphism exists. However, whilst the expansion is considerably more than many mammalian species, the homology between each of these proteins is not comparable to that of the MUPs. The lack of genome data for *T. vulpecula* means that it is not possible to tell if this potential polymorphism also exists in the brushtail possum from genome data, and where it might be expressed, if so. The only evidence for closely related protein sequences is that of trichosurin, expressed during lactation of *T. vulpecula*.

There is some suggestion that trichosurin is glycosylated; the same disparity is observed between SDS-PAGE resolved molecular weight and the size of the sequenced peptide (Piotte *et al.*, 1998), and automatic assignment in UniProt indicates two putative N-glycosylation sites at residues 67 and 148 (Bateman *et al.*, 2017). The sequence of the two β-lactoglobulins identified at the same time as trichosurin also don't correlate with their resolved molecular masses, suggesting a post-translational modification (Piotte *et al.*, 1998). The identification of MUP-like lipocalins in urine and milk, both important in chemosignalling within the animal kingdom, suggest that this considerable investment in

protein output may have a role in chemosignalling. It also highlights the potentially important role of glycosylation, which, as post-translational modifications are not template-driven, can be difficult to compare between and within species on a large scale.

Another aspect not considered was the ligand binding capability of vulpeculin. As the majority of samples were freeze dried for transportation, the low molecular weight component of urine samples could not be investigated. Worth revisiting is that scent signalling behaviours in the possum also exhibit use of sternal and glandular secretions, and these may also involve the use of lipocalin sequences. Overall, the identification and sequencing of vulpeculin opens up further questions into chemosignalling mechanisms in marsupials, and highlights the large knowledge gap in marsupial biology.

This knowledge gap is exemplified when attempting discovery proteomics using cross-species matching. The quality of protein identifications was not at a confident level. Whilst proteomic analysis of urine confidently identified approximately 100-200 proteins in the other species explored within this thesis, analysis of *T.vulpecula* urine by cross-species matching to unannotated genome data from *M.domestica* only identified 57 proteins, 16 of which were of high enough quality for label-free quantification. This highlights the deficit of marsupial research, perhaps because of their unique evolutionary lineage.

The brushtail possum, *T.vulpecula*, is the subject of many conservation efforts in New Zealand, where it is a non-native species that inhibits survival of native flora and fauna. A protein has been identified and sequenced in both male and female possum urine, which with further research may help to progress pest control efforts.

5.6 Characterisation of the urinary protein content in the New Zealand brushtail possum, *Trichosurus vulpecula:* Supplementary information

5.6.1 Intact mass profiles of individual samples

5.6.2 MALDI-ToF peptide mass fingerprints from tryptic in-gel digestion

5.6.3 Homologous sequences to the trichosurin-like protein from M. domestica

5.6.4 Fragment ion spectra for sequencing

5.6.5 Fragment ion spectra for determination of leucine and isoleucine residues

5.6.6 Table of accession numbers for multiple sequence alignment

5.6.7 Multiple sequence alignment of lipocalins

6 General Discussion

6.1 Summary of results and implications in behaviour and social structure

The main aim of this thesis was to use proteomics techniques to explore the urinary proteomes of mammalian species, beyond that of the well-characterised house mouse and Norway rat. Proteins within male and female urine of the bank vole (*Myodes glareolus*), field vole (*Microtus agrestis*) and brushtail possum (*Trichosurus vulpecula*) were investigated, in addition to the scent marks of both vole species (Table 6.1).

Urinary expression of the major urinary proteins of the house mouse (*M. musculus*) is polymorphic, and is used to communicate individual identity, mating availability, hierarchy and kinship. This complex olfactory mechanism has been associated with the dense social structure of the house mouse; the populated habitats in which it resides increase the need for hierarchical structuring, and results in a higher chance of interaction with a related individual, leading to a necessity to avoid inbreeding in this polygamous species. Protein heterogeneity is also observed in rat (*R. norvegicus*) urine. A large multigene family of MUPs is located on chromosome 5, and although social and reproductive behaviours are less well understood than in the house mouse, urine marking is hormonally-controlled and can transmit information including age, sex, reproductive status, individual identity and dominance (Brown, 1988, 1995; Gómez-Baena *et al.*, 2014, 2019). Both the Norway rat and the house mouse exist in densely populated habitats, and a complex olfactory communication is likely an essential component of establishing social structures. However, the majority of mammalian species don't live in the same dense populations, and olfactory mechanisms underlying reproductive success and social hierarchy are under alternative pressures. Even within the *murid* family of rodents, MUP expression differs. The steppe mouse (*Mus spicelegus*), although genetically close to the house mouse, is a monogamous species with at least four functional MUP gene products, three of which are male-specific (Lee, 2015). Urine from male members of *Mus macedonicus*, an aboriginal grassland species that lives independently of humans, is uniform, containing a single MUP that is closely related to those of the house mouse (Robertson *et al.*, 2007). Likewise, male urine of the species *Mus spretus* contains just three MUPs (Beynon *et al.*, 2008).

However, MUPs are not the only lipocalins in mammalian scent marking. In *cricetid* rodents, odorant binding proteins are more commonly observed in scent secretions than MUPs. For example, urine and saliva of *Phodopus sungora* contains at least one OBP, resolving in three

bands on SDS-PAGE, although sexual dimorphism is not known. Urine of both sexes of *Phodopus roborovskii* contains a single OBP (Turton *et al.*, 2010). The water vole, *Arvicola amphibius*, contains a sexually dimorphic protein profile, with protein bands resolving at molecular weights indicative of a lipocalin (Nazarova, Proskurniak and Yuzhik, 2016). Aphrodisin, male-specific secretory protein (MSP) and female-specific extraorbital lacrimal gland protein (FLP) are expressed in the golden hamster, *M. auratus*, in vaginal fluid, the submandibular salivary gland and the extraorbital lacrimal glands, respectively (Ranganathan and De, 1995; Briand, Trotier and Pernollet, 2004; Dubey *et al.*, 2013). OBPs are also expressed with a small degree of polymorphism in the nasal tissue of the porpupine and the giant panda (Felicioli *et al.*, 1993; Zhu *et al.*, 2017), in addition to single variant expression in rabbit seminal fluid (Mastrogiacomo *et al.*, 2014), suggesting lipocalin secretion is not limited to rodents, and that lipocalins may possibly play a role in chemosignalling across the mammalian kingdom.

However, it is worth bearing in mind that lipocalins are not the only proteins with roles in chemosignalling. One example is cauxin, an esterase in the pathway that produces felid-specific felinine, which decomposes to 3-mercapto-3-methyl-1-butanol (MMB), a male sex recognition pheromone (Miyazaki *et al.*, 2018). Another is WFD12, a whey acidic protein secreted into the urine of a subset of male mouse lemurs during the breeding season at hugely increased levels compared to outside the breeding season, or to males not belonging to the subset and females (Unsworth *et al.*, 2017) (joint first author, supplementary material). Both enzymes have the potential to orchestrate chemical signals within respective scent secretions, and the metabolic expense of excreting large quantities of protein suggests this is a highly likely explanation for their presence. These examples open up the possibility that proteins with roles in chemosignalling are not limited to lipocalins, and that the chemical profile of scent secretions can be influenced by the protein content in a variety of mechanisms.

6.1.1 Myodes glareolus

It is clear that protein expression in scent secretions differs vastly between species, in terms of native polymorphism and sexual dimorphism. In chapter 3, the protein content of bank vole urine was explored. A single, seasonally-expressed, male-specific protein (glareosin) was identified and sequenced fully, with homology to previously reported bank vole OBP sequences (Stopková *et al.*, 2010a). This work was published (Loxley *et al.*, 2017), but work exclusive to this thesis explored male urinary proteome complexity at lower abundances than the predominant protein glareosin, identifying peptide-level evidence of OBP-like

heterogeneity, including those previously reported. However, the dominant protein in breeding season male urine is unequivocally glareosin in almost all cases, and other quantifiable OBP proteins were generally at similar or lower abundance to serum albumin (Figure 3.1), and may be indicative that these proteins are expressed for function elsewhere, and simply 'leak' into the urine. Alternatively, they could contribute to the overall chemosignalling profile (either individually or via bound ligands), finely tuning the information within the scent mark despite the low abundance. Indeed, a couple of individual samples revealed higher levels of OBP3 and OBP1. Unfortunately discovery proteomics analysis was not conducted on out-of-breeding season samples so we do not yet know if this low level expression of OBP-like proteins is also seasonal. Whilst it is still unclear what role these proteins play, glareosin's seasonal expression strongly suggests a function in chemosignalling. Investigation into the protein content of bank vole scent marks also suggests simple protein content, although an alternative, MS-friendly sampling technique would benefit further research to enable intact mass profiling of a larger sample number. The higher protein content of female scent marks in comparison to urine suggests a distinctive difference in function between the two fluids, with OBP1 identified particularly strongly in scent marks. In contrast, male scent marks contained similar proteins to that in urine; glareosin was still the most confidently identified protein, although strong coverage of OBP3 was also observed. In males, distinction between urine intended for excretion and for scent marking is difficult; as opposed to females, which void urine in pools, males deposit urine in smaller volumes, streaking out the mark with the prepuce tip (Johnson, 1975; Christiansen, 1980). The method of urine recovery confines the vole on an elevated mesh in a new cage, so that deposited urine is able to be collected from underneath. However, this may prevent the vole from 'streaking' the urine, indicative of a scent mark, therefore interfering with distinction between the two secretions, if there is any. Additionally, males increase scent marking rates in new environments (Johnson, 1975). Both male and female scent marks however also contained other OBP-like sequences that differ between the sexes, and indicates heterogeneity.

Overall, the protein profiles of both male urine, and male and female scent marks, seem generally simple, and are supportive of the observed correlation between species population density and chemosignalling mechanism complexity; bank voles are generally solitary during the breeding season, with a small overlap between male home ranges, and scent marking behaviour contributes to hierarchical structure and mating opportunities (Mazurkiewicz, 1981; Bujalska, 1990; Kruczek, 1997; Kruczek and Gołas, 2003). However,

given the level of heterogeneity at low protein abundances, from identification of established OBPs to manually identified homologous sequences, the complexity of scent marking in bank voles could be more complex than first thought. The domination of the protein profile by glareosin makes identification of these lower abundance OBP proteins difficult, however. Without genome data, understanding the total possible heterogeneity of odorant binding proteins in bank vole urine and scent marks is challenging.

6.1.2 Microtus agrestis

Analysis of field vole urine and scent marks revealed a different protein profile than that of the bank vole. Lipocalin content of both male and female urine samples was more complex than observed in the bank vole, and likewise the field vole scent mark samples were also more complex with regard to identified lipocalin species than observed for the bank vole. Proteins closely related to bank vole glareosin were identified in mature male urine of field voles. At least three variants of field vole glareosin were identified and sequenced, and a fourth major mass peak observed on intact protein mass analysis is a result of another possible mutation. The smallest variant has a total mass of 17143 Da. Another variant of mass 17172 Da and of the 149 amino acids, was the result of a single point mutation 97[A>V] with respect to the smallest variant. The variant of mass 17240 Da was the result of seven point mutations with respect to the smallest variant, of 85[L/I>M], 89[L/I>F], 97[A>V], 113[D>G], 115[H>P], 118[P>T] and 121[S>Y]. One other potential variant was also observed, the sequence of which, accounting for disulfide bond formation in the intact protein, would have an average mass of 17256 Da. This could either be the result of an oxidation modification, or the single point mutation 37[N>K], with respect to the variant of mass 17240 Da.

According to structural homology modelling, all point mutations across the proteins were located on the outside of the beta-barrel, suggesting diversity in external interactions rather than changing potential ligand-binding properties of the internal cavity. It is possible that, if field vole glareosin were to act directly as a pheromone, in a similar way to darcin in the house mouse, this surface-level heterogeneity is important in detection and the associated response. Investigation into the volatile components of field vole urine, and specifically, field vole glareosin, in addition to behavioural responses, was beyond the scope of this thesis. Field vole glareosin is male-specific, identified in no female urine or scent marks, and the four contributing mass peaks dominate protein profiles, similar to expression of bank vole glareosin. However, in contrast to bank voles, protein output in female urine of field voles is considerable. Whilst still significantly less than male protein

output, female urine contains protein bands visible on SDS-PAGE beyond that of the typical output of albumin, uromodulin and other higher molecular weight serum proteins. The main protein, or proteins, in female field vole urine was discovered to be MUP-like, with multiple mass peaks visible from intact protein mass analysis at approximately 21 kDa. A draft sequence of overlapping peptides sequenced *de novo* from MS/MS analysis of female field vole urine, was constructed, bar a missing portion of sequence near the N-terminal, which is proposed to contain a glycosylation site, for reasons discussed in chapter 4. Both MUPs and OBPs expressed in tissues and secretions involved in olfaction have been reported as being glycosylated. Boar salivary lipocalin (Loebel *et al.*, 2000), OBP3 of boar nasal epithelia (Scaloni *et al.*, 2001) and MUP3 of the house mouse (Mechref *et al.*, 2000) are all examples, although any role of the attached glycan in chemosignalling is yet to be established. However, given the importance of the above proteins in chemosignalling of their respective species, glycan moieties may be of significance.

The discovered MUP-like protein, or proteins, was also prevalent in male field vole urine; intact protein analysis was dominated by field vole glareosin, but the MUP-like protein was still visible as 21 kDa masses. Of interest, LFQ of proteins identified in mature field vole urine indicates the sequenced MUP is the most abundant protein, in males and females. However, quantification of glareosin variants is complicated by the near-identical sequences. Furthermore, additional OBP sequences, including cross-species identified matches to bank vole OBPs and an epididymal-specific lipocalin, were found from manual inspection of unidentified peptides. Similarly to the bank vole, this was indicative of additional unknown complexity. Following analysis of juvenile urine samples and scent marking samples, and subsequent re-assessment of the mature field vole urine, peptide evidence of a partially sequenced OBP3-like protein was observed, in addition to another lipocalin denoted here lipocalin 11. Although the expression of field vole glareosin is in many ways similar to bank vole glareosin expression, the urinary proteome of the mature field vole is far more complex than that of the bank vole. A small number of polymorphic glareosin variants was observed, but the most notable complexity is driven by the number of different lipocalin types observed. Previous investigations suggest that cricetid rodents have a tendency to express OBPs rather than MUPs in scent secretions, however urine from mature field voles of both sex express both, and potentially more. Rather than considering MUPs and OBPs the 'equivalent' in terms of function in each rodent family, this could suggest that there are distinctive and specialised roles for each.

The opportunity also arose to initiate a preliminary investigation into the urinary protein content of juvenile field voles. Scent signalling is usually associated with sexual selection, either in attraction of a mate or by establishing dominance to gain access to mating opportunities. Juvenile mammals can detect odour signals, even acting to accelerate or delay puberty onset in mice (Massey *et al.*, 1980; Novotny *et al.*, 1986; Novotny *et al.*, 1999), however they are also capable of using protein-mediated signalling, for example to prevent mating behaviour of sexually mature individuals (Ferrero *et al.*, 2013). Both male and female juvenile field voles express urinary protein. The presence of expressed protein, even at a lower level than adult males, is indicative of a role in scent signalling. Whilst sample numbers were not high enough to form concrete conclusions or observable patterns, consistent peptide-level evidence of the novel MUP-like sequence was notable, in addition to the constructed OBP3-like protein sequence. Male juvenile samples also contained various levels of field vole glareosin. Urinary protein expression in juvenile field voles is clearly heterogeneous and complex. Studies with a greater sample number could investigate further into sex-specific differences, protein expression over time, inter- and intra-litter variation, and the associated behavioural responses, either on siblings, parents or unrelated conspecifics. Despite the unresolved complexity, it is highly likely that juvenile field voles utilise urine markings in olfactory communication.

A similar story was revealed after proteomics examination of field vole scent marks. Similar to the bank voles, intact mass analysis was not possible, so assessment of overall protein heterogeneity is limited to PAGE. Nevertheless, SDS-PAGE separated a number of protein bands in the 17 kDa and 6.5 kDa regions of both male and female scent mark samples. In both males and females the 6.5 kDa bands were identified, with varying confidence, as secretoglobin. However, in the 17 kDa region, peptide evidence from female samples most consistently suggested an OBP3-like protein, whereas gel band identification for male samples in this region was far more difficult. Manual inspection of *de novo* peptide candidates of high quality spectra suggested heterogeneous OBP-like proteins, one of which was most closely related to a protein type denoted here lipocalin 11, but with conflicting terminology in associated database entries. Whilst again, proteomics investigation of scent mark secretion has thus far offered no concrete conclusions, it has demonstrated the expression of protein in scent marks, and offered peptide-level evidence of another type of lipocalin present in mammalian scent secretions. It has also suggested high levels of polymorphism in OBP-like proteins. The true extent of this however is unknown. Genome analysis would be an ideal step forward for investigating the potential

for polymorphism, in addition to providing a reference genome for proteomics database searching, and profiling of a larger number of samples may give a better idea of the expressed polymorphism.

There is very little known regarding the social behaviours and communication of field voles. Vocalisations are employed more frequently than in bank voles, and are important in sexual selection (Mandelli and Sales, 2004; Sales, Czuchnowski and Kapusta, 2007; Kapusta and Sales, 2009), however no investigation has been performed on olfactory communication of field voles. A predominantly solitary species, interactions are mainly during the breeding season between males in nearby home ranges or for copulation (Breed and Clarke, 1970; Agrell *et al.*, 1996). Considering the field vole is mainly a solitary species, it is difficult to combine the observed protein complexity with the theory that it is reflective of social structure. However, social complexity of the field vole is apparent instead in population cyclicity; populations fluctuate over a period of four years, which is even correlated with changes from nocturnal activity to diurnal (Halle and Lehmann, 1992; Norrdahl, 1995). A long-term investigation into the communication of field voles over this period, both auditory and olfactory, may give further insight into this unexplained phenomenon.

6.1.3 Trichosurus vulpecula

Similarly to the field vole, the brushtail possum is generally solitary with small, overlapping home ranges (Crawley, 1973; Ward, 1984). Social interactions rarely occur outside of the breeding season, and scent marking behaviours are exhibited by dominant individuals, particularly males (Spurr and Jolly, 1999; Ji, White and Clout, 2005; McLean, 2014). A couple of findings hint at the use of protein components in scent communication of marsupials; firstly, they have a well-developed vomeronasal organ similar to rodents (Poran, 1998; Schneider *et al.*, 2008; Aland, Gosden and Bradley, 2016), and second, lipocalins are expressed in the milk of a number of marsupials, a possible source of olfactory communication (McKenzie, Muller and Treacy, 1983; Nicholas *et al.*, 1987; Collet, Joseph and Nicholas, 1989; Beg and Shaw, 1994; Piotte *et al.*, 1998; Trott *et al.*, 2002). Considering the use of urine to transmit scent marks, it followed that the brushtail possum could exhibit urinary protein expression. A single, glycosylated protein was found in urine from both males and females, and sequenced completely, including discrimination between isobaric amino acids leucine and isoleucine, using a dietary isotope labelling approach identical to that taken for bank vole glareosin. Sequence analysis and homology modelling suggested a close structural similarity to the major urinary proteins of the house mouse and rat; a single conserved disulfide bond is likely to promote structural stability. However, phylogenetic

analysis grouped the completed sequence in a clade with other lipocalin sequences from unannotated marsupial genomes, separate to other established lipocalin sub-families within placental mammals. Within the marsupial lipocalins, three further clades can be distinguished, one of which forms a distantly related clade with the outlier OBPs. The largest group of marsupial lipocalins forms an extended clade with the first, and another is grouped separately, although still more closely related to other marsupial lipocalins than the main placental mammal lipocalin sub-families. This indicates that members of the lipocalin family were likely present before the division of *Metatheria* and *Eutheria*, however the distinctive sub-families established in placental mammals did not arise until after this event, with the possible exception of a subset of odorant binding proteins. This early speciation event, combined with the knowledge that marsupials express lipocalins in potential sources of chemical information such as milk and urine, just like in many placental mammals, could provide information on the evolution of chemical signalling mechanisms. It is possible that these barrel-shaped proteins have had an integral role in semiochemical pathways since the speciation event. However, without further evidence of a behavioural response to the protein, we cannot be sure that these proteins are utilised in the same way as in placental mammals.

Furthermore, this investigation was a preliminary probe into the protein content of brushtail possum urine. In terms of the protein chemistry, we are unaware of the structure of the attached glycan, and any potential heterogeneity or sexual dimorphism. It is also difficult to rule out the possibility that higher levels of heterogeneity exists at the protein level that was not detected by intact protein analysis due to the interference of the glycan; even the spectra of deglycosylated protein were not as abundant as could be desired as a result of the aggressive sample preparation required for the deglycosylation of this structurally stable protein. Additionally, the particularly difficult identification by cross-species matching of a marsupial, discussed in chapter 5, may mask cross-species matching of lower abundant proteins to other lipocalin species, as was observed in the bank vole and field vole. Another future direction in chemical analysis could examine the possibility of bound ligands, which was not considered in this analysis.

Vulpeculin is expressed in the urine of both sexes. It is therefore likely that it is not used in the attraction of one sex to another, and a role in establishing dominance is perhaps a more plausible explanation. The possibility of sexual dimorphism or individual differences in the attached glycan needs to be investigated further, however, as does the existence of a bound ligand, and their associated effects on behavioural response. Another aspect not

investigated was any change of protein expression with seasonality and maturity. Relative to the theory that signalling profile is a reflection of social structure complexity, the simple protein expression thus far established fits well. Possums are predominantly solitary animals, and given the hypothesised link between chemosignalling protein complexity and social structure, the expression of a single main protein species is not unexpected. However, this ignores potential heterogeneity and sexual dimorphism in the attached glycan.

The discovery of urinary and scent secretion lipocalins, related to those with a known role in chemosignalling, in such a diverse selection of species, is representative of a huge knowledge gap. However, characterisation of proteins under evolutionary pressure remains a challenge, particularly for species without a reference genome.

Table 6.1 | Summary of chapter findings.

Common name	Species	Sample type	Sexual dimorphism	Polymorphism	Main findings
Bank vole	*Myodes glareolus*	Urine	✓	X	• Single dominant protein (glareosin) in breeding season males. • Females very little protein expression • Low abundance peptide-level evidence for previously established bank vole OBPs, and additional OBP-like sequences in male urine
		Scent marks	✓	X	• High levels of protein expression in both male and females • Evidence of OBPs and homologous OBP-like sequences (different between males and females, and different to OBP-like sequences in male urine)
Field vole	*Microtus agrestis*	Urine	✓	✓	• Field vole glareosins, OBP-like proteins expressed exclusively in male urine, were the dominant proteins in samples from mature individuals. • At least 3 variants of glareosin are expressed at high abundance in mature male urine. • Peptide evidence for a major urinary protein-like lipocalin, an OBP3-like protein and a lipocalin 11-like protein identified in all samples; in both male and female samples, and in urine from both mature and juvenile individuals. • Peptide evidence for OBP-like protein heterogeneity at low levels in the urine of both sexes and maturity stages.
		Scent marks	✓	✓	• Sexually dimorphic expression of glareosin • Strong evidence for the OBP3-like protein, lipocalin 11-like, and to some extent major urinary protein-like. • No peptide-level evidence of bank vole OBP1 or epididymal-specific lipocalin-like proteins.
Brushtail possum	*Trichosurus vulpecula*	Urine	X	X	• Single lipocalin protein expressed in both male and female urine, named vulpeculin. • Attached glycan of approximately 2 kDa, the structure of which is potentially sexually dimorphic and/or polymorphic. • Vulpeculin grouped within a novel class of marsupial lipocalins.

6.2 Methodology

Proteomics on non-model organisms remains a challenge, and whilst various tools are disposed to tackle the unique challenges of cross-species identification, there are important caveats. Characterisation of proteins under acutely high evolutionary pressure, such as those involved in chemosignalling, is particularly challenging. The approach taken for the characterisation of proteins in the urine and scent marks of bank voles, field voles and brushtail possums took into careful consideration the limitations of cross-species proteomics, and highlighted a number of problematic areas.

Identification by cross-species matching relies on good quality matches between the experimentally investigated species and the reference database. When investigating proteins that vary in levels of polymorphism, even between closely related species, this can become an issue. For example, the gene duplication of MUPs in the house mouse results in a large number of potential protein identifications when used as a cross-species database. If the sample investigated contained peptide matches to different members of this family but in exclusive portions of the sequences, it could not be determined if the investigated sample contained peptides from different MUP-like species, or from a single MUP-like species that coincidentally matched different regions of the database proteins. It is therefore useful to investigate intact protein profiles to explore overall complexity, as performed here. However, it is important to bear in mind that, as demonstrated during analysis of both field and bank vole scent secretions, lower abundance proteins still contribute to the peptide landscape and requires careful manual consideration. Additionally, post-translational modifications can affect intact mass profiles, as observed during analysis of vulpeculin and field vole MUP, and whilst this can be instrumental in exploring post-translational complexity, it is detrimental to understanding heterogeneity of the polypeptide complement only.

In some cases, it is worth identifying a pure, or low complexity sample, for example, by digestion of proteins separated by PAGE coupled with LC-MS/MS analysis. This approach reduces the amount of peptide interference from proteins outside those of interest, in terms of both mass spectrometric analysis and data searching. A focused approach highlighted differences in identification confidence between conserved housekeeping proteins, such as albumin and uromodulin, and those proteins likely to be under evolutionary pressure. In all species investigated, housekeeping proteins were the most confidently identified by cross-species matching. The approach was less successful for

identification of lipocalin sequences. In particular, in-gel identification of male field vole scent mark proteins and brushtail possum urinary proteins were not confidently identified and required manual inspection of unmatched spectra, a labour-intensive approach. In-gel identification, whilst supposedly a simple 'first-pass' look, is limited in peptide recovery and protein resolution. For higher-throughput identification from multiple samples, 1D PAGE is usually employed, which would struggle to separate closely related proteins to a degree of visible separation. 2D gels are an alternative, separating proteins by both size and charge, but are far more labour intensive.

An alternative approach taken here was to isolate proteins of interest using chromatographic separation. A common technique for protein purification, it is possible to purify proteins at a higher resolution than possible with 1D PAGE, and has the added bonus of recovering the intact protein, therefore allowing monitoring of the intact protein using ESI-MS to determine the eluted protein profiles and in-solution digestion, which is preferable to in-gel proteolysis in terms of digestion completeness. Whilst more information rich, chromatographic purification is labour intensive, particularly if optimisation steps are required, and can require large amounts of protein with little return, an issue with limited sample volumes or numbers. However, it does effectively separate proteins for sequencing *de novo*, for use in behavioural assays, or for investigating the volatile ligands associated with each protein species.

Global proteomics, including identification and quantification, gives an opportunity to address the total protein content of a sample. Global proteomics gives an opportunity to identify proteins other than those that dominate intact mass spectra or SDS-PAGE, giving another dimension of protein complexity at the peptide level. For example, in both vole species, peptide-level heterogeneity was discovered from manual inspection of peptides sequenced *de novo*, and from *de novo* tags mapped to identified proteins. In the analysis of scent secretions, other proteins within a sample can indicate the tissue of origin. Prostate gland-specific proteins, for example, may indicate that part of the urinary protein content originates from a glandular secretion, including those proteins whose expression profiles correlate, and therefore, quantification of proteins can add depth of understanding. Furthermore, differences in abundance between scent signalling proteins can be functionally relevant; the unique ratio of MUPs in house mouse urine is thought to convey individual identity (Roberts *et al.*, 2018). Simple determination of presence or absence of a protein may not be enough to fully understand differences that shape these odour profiles. However, quantification of proteins identified by cross species matching presents additional

challenges. Labelled approaches of specific proteins, such as the use of QconCATs (Pratt *et al.*, 2006), can be ideal for accurate and reliable quantification of already established protein targets. However, for a first-pass assessment of protein abundance, in order to ascertain patterns, labelled approaches are costly, both in time and budget. Likewise, a targeted MS approach is dependent on reliable peptides from which to assay quantifiable transitions, and similar to labelled approaches would be a suitable solution to established target proteins. LFQ is the most accessible technique and is wholly more suitable to preliminary assessments due to the low sample preparation cost and labour, although it is generally more demanding on MS time. Quantification using the top three most intense, high-quality peptides is the most common approach (high N), however, the number of peptides identified by cross-species matching is considerably lower than matching an annotated genome of the analysed species, therefore reducing the number of quantifiable proteins. Additionally, the possibility of incorrect identification is raised, therefore reducing the reliability of the data. With respect to the scent signalling proteins discussed in this thesis, it is also possible for the number of unique peptides to be compromised due to the sequence similarity of some lipocalin families. For example, some house mouse MUPs only differ by one amino acid, preventing LFQ using high N approaches. This was problematic in quantification of field vole glareosin, in which one variant had no unique tryptic peptides- every peptide was shared with at least one other variant of field vole glareosin. With prior sequence knowledge, this could potentially be circumvented by the use of alternative proteolytic enzymes, however this is reliant on good, specific digestion efficiency suitable for the protein, or proteins, in question. An alternative option is to observe patterns in peptide features, rather than identified peptides, by identifying MS/MS features that contribute to the most significant abundance changes between experimental groups and use the unidentified features as a starting point from which to identify proteins. However, this laborious approach may still struggle to identify proteins, and therefore attribute functional significance to findings. In all samples investigated in this thesis, quantification was only achieved on a small number of proteins. Whilst quantification of housekeeping proteins identified by cross-species matching may achieve better results (for example, albumin was confidently quantified each time), cross-species identification of proteins under higher evolutionary pressure is too flawed for quantification using this approach. Furthermore, successful cross-species matching is conditional on the quality and relatedness of the database. Whilst choosing a database organism closest to that of the experimental data can increase the likelihood of a PSM, it may be limited by the database quality. Inconsistency in genome annotation was emphasised during phylogenetic analysis

of marsupial lipocalins from entries generated by gene prediction modelling of genome data. Many marsupial lipocalins were denoted 'MUP-like' or 'allergen-like', whereas phylogenetic analysis of these proteins revealed that the marsupial lipocalins formed a clade of their own, rather than grouping with established lipocalin classes of placental mammals, such as MUPs or allergen proteins. This provides an example of how using an uncurated database could lead to identification of a protein sequence that, labelled incorrectly, could imply false functionality. However, in this case, marsupial genetics are so far removed from the few species with sequenced genomes that using a poorer quality database with more similarity to the subject was beneficial for identification and limited quantification. An additional challenge to address is the choice between a single species database or a database with multiple species entries. Searching against a database with multiple species entries can increase the likelihood of a PSM, however the number of matches against the same protein of different species can cause problems in quantification and identification validation.

With the advent of faster, cheaper and more efficient genome sequencing capabilities, future directions will benefit from complete reference databases before attempting discovery proteomics and LFQ. However, care should always be taken in identification and quantification of protein samples based on genomic data; a truncated version of a protein, the result of an alternative intron/exon sequence, was discovered in male mouse lemur urine (Unsworth *et al.*, 2017).

6.3 Summary

Concerted efforts have resulted in a firm understanding of protein-mediated scent communication in species such as the house mouse and the rat, however information on how this mechanism transcends mammalian lineages is lacking. Presented above is the characterisation of protein output in three different mammalian species. Evidence of urinary and scent mark protein output, particularly members of the lipocalin family, in *M. glareolus* and *M. agrestis*, demonstrates that protein expression, whilst clearly still important among other rodent species, does not always follow the same polymorphism and sexual dimorphism that is displayed in mice and rats. Both species are indicators of alternative expression profiles, the complexity of which are suggested to correlate with social structure and population density. Evidence of a lipocalin expressed in the urine of a marsupial, *T. vulpecula*, emphasises the importance of this protein expression across divergent mammalian taxa.

Considerable further work is required, both in terms of chemical characterisation and behavioural significance, to understand fully the implications of these findings. However, this work serves as a base from which to form new questions and design future experiments, in addition to highlighting the complications that arise with proteomic analysis of species without a reference genome. It has also highlighted the large gap in scientific knowledge of olfactory communication, which in favour of auditory and visual communication has been largely ignored, but is evidently of immense importance across the animal kingdom.

7 References

Abdul Rahman, S. *et al.* (2014) 'Filter-aided N-glycan separation (FANGS): A convenient sample preparation method for mass spectrometric N-glycan profiling', *Journal of Proteome Research*. American Chemical Society, 13(3), pp. 1167–1176. doi: 10.1021/pr401043r.

Achiraman, S. *et al.* (2010) '1–Iodo-2 methylundecane [1I2MU]: An estrogen-dependent urinary sex pheromone of female mice', *Theriogenology*, 74, pp. 345–353. doi: 10.1016/j.theriogenology.2010.01.027.

Agrell, J. (1995) 'A shift in female social organization independent of relatedness: an experimental study on the field vole (*Microtus agrestis*)', *Behavioral Ecology*. Narnia, 6(2), pp. 182–191. doi: 10.1093/beheco/6.2.182.

Agrell, J. *et al.* (1996) 'Shifting spacing behaviour of male field voles (*Microtus agrestis*) over the reproductive season', *Annales Zoologici Fennici*. Finnish Zoological and Botanical Publishing Board, 33, pp. 243–248. doi: 10.2307/23735780.

Aland, R. C., Gosden, E. and Bradley, A. J. (2016) 'Seasonal morphometry of the vomeronasal organ in the marsupial mouse, *Antechinus subtropicus*', *Journal of Morphology*, 277(11), pp. 1517–1530. doi: 10.1002/jmor.20593.

Allen, R. B., Fitzgerald, A. E. and Efford, M. G. (1997) 'Long-term changes and seasonal patterns in possum (Trichosurus vulpecula) leaf diet, Orongorongo Valley, Wellington, New Zealand', *New Zealand Journal of Ecology*, 21(2), pp. 181–186.

Allison, A. C. and Rees, W. A. (1957) 'The binding of haemoglobin by plasma proteins (haptoglobins); its bearing on the renal threshold for haemoglobin and the aetiology of haemoglobinuria.', *British medical journal*. BMJ Publishing Group, 2(5054), pp. 1137–43. doi: 10.1136/bmj.2.5054.1137.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*. Academic Press, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Anderson, N. L. and Anderson, N. G. (2002) 'The Human Plasma Proteome History, Character, and Diagnostic Prospects', *Molecular & Cellular Proteomics*, 1(11), pp. 845–867. doi: 10.1074/mcp.R200007-MCP200.

Apps, P. J. (2013) 'Are mammal olfactory signals hiding right under our noses?', *Naturwissenschaften*. Springer-Verlag, 100(6), pp. 487–506. doi: 10.1007/s00114-013-1054-1.

Apweiler, R. *et al.* (2004) 'UniProt: the Universal Protein knowledgebase', *Nucleic Acids Research*, 32, pp. D115–D119. doi: 10.1093/nar/gkh131.

Archie, E. A. and Theis, K. R. (2011) 'Animal behaviour meets microbial ecology', *Animal Behaviour*. Academic Press, 82(3), pp. 425–436. doi: 10.1016/J.ANBEHAV.2011.05.029.

Arike, L. *et al.* (2012) 'Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli', *Journal of Proteomics*. Elsevier, 75(17), pp. 5437–5448. doi: 10.1016/J.JPROT.2012.06.020.

Armstrong, S. D. *et al.* (2005) 'Structural and functional differences in isoforms of mouse major urinary proteins: a male-specific protein that preferentially binds a male pheromone.', *The Biochemical journal*. PORTLAND PRESS LTD, THIRD FLOOR, EAGLE HOUSE, 16 PROCTER STREET, LONDON WC1V 6 NX, ENGLAND, 391(Pt 2), pp. 343–50. doi: 10.1042/BJ20050404.

Arteaga, L. *et al.* (2013) 'Smell, Suck, Survive: Chemical Signals and Suckling in the Rabbit, Cat, and Dog', in *Chemical Signals in Vertebrates 12*. New York, NY: Springer New York, pp. 51–59. doi: 10.1007/978-1-4614-5927-9_4.

Bacchini, A., Gaetani, E. and Cavaggioni, A. (1992) 'Pheromone binding proteins of the mouse, Mus musculus', *Experientia*. BIRKHAUSER VERLAG AG, PO BOX 133 KLOSTERBERG 23, CH-4010 BASEL, SWITZERLAND, 48(4), pp. 419–421. doi: 10.1007/BF01923448.

Bachmann, S. *et al.* (2005) 'Renal effects of Tamm-Horsfall protein (uromodulin) deficiency in mice', *American Journal of Physiology-Renal Physiology*. American Physiological Society, 288(3), pp. F559–F567. doi: 10.1152/ajprenal.00143.2004.

Bachmann, S., Koeppen-Hagemann, I. and Kriz, W. (1985) 'Ultrastructural localization of Tamm-Horsfall glycoprotein (THP) in rat kidney as revealed by protein A-gold immunocytochemistry.',

Histochemistry, 83(6), pp. 531–8.

Bachmann, S., Metzger, R. and Bunnemann, B. (1990) 'Tamm-Horsfall protein-mRNA synthesis is localized to the thick ascending limb of Henle's loop in rat kidney.', *Histochemistry*, 94(5), pp. 517–23.

Bantscheff, M. *et al.* (2012) 'Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present', *Analytical and Bioanalytical Chemistry*. Springer-Verlag, 404(4), pp. 939–965. doi: 10.1007/s00216-012-6203-4.

Barlow, N. D. (1994) 'Predicting the Effect of a Novel Vertebrate Biocontrol Agent: A Model for Viral-Vectored Immunocontraception of New Zealand Possums', *The Journal of Applied Ecology*. British Ecological Society, 31(3), p. 454. doi: 10.2307/2404442.

Bastian, F. *et al.* (2008) 'Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species', in *Data Integration in the Life Sciences*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 124–131. doi: 10.1007/978-3-540-69828-9_12.

Bateman, A. *et al.* (2017) 'UniProt: the universal protein knowledgebase', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Bates, D. *et al.* (2015) 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. doi: 10.18637/jss.v067.i01.

Baumgartner, C. *et al.* (2008) 'SeMoP: A New Computational Strategy for the Unrestricted Search for Modified Peptides Using LC−MS/MS Data', *Journal of Proteome Research*. American Chemical Society, 7(9), pp. 4199–4208. doi: 10.1021/pr800277y.

Bautista, A. *et al.* (2005) 'Scramble competition in newborn domestic rabbits for an unusually restricted milk supply', *Animal Behaviour*. Academic Press, 70(5), pp. 1011–1021. doi: 10.1016/J.ANBEHAV.2005.01.015.

Bautista, A. *et al.* (2008) 'Do newborn domestic rabbits *Oryctolagus cuniculus* compete for thermally advantageous positions in the litter huddle?', *Behavioral Ecology and Sociobiology*. Springer-Verlag, 62(3), pp. 331–339. doi: 10.1007/s00265-007-0420-4.

Bayram, H. L. *et al.* (2016) 'Cross-species proteomics in analysis of mammalian sperm proteins', *Journal of Proteomics*. Elsevier, 135, pp. 38–50. doi: 10.1016/J.JPROT.2015.12.027.

Beg, O. U. and Shaw, D. C. (1994) 'The complete primary structure of late lactation protein from quokka (*Setonix brachyurus*)', *Journal of Protein Chemistry*, 13(6), pp. 513–516. doi: 10.1007/BF01901532.

Benkert, P., Tosatto, Silvio C E and Schomburg, D. (2008) 'QMEAN: A comprehensive scoring function for model quality assessment.', *Proteins*, 71(1), pp. 261–77. doi: 10.1002/prot.21715.

Benkert, P., Tosatto, Silvio C. E. and Schomburg, D. (2008) 'QMEAN: A comprehensive scoring function for model quality assessment', *Proteins: Structure, Function, and Bioinformatics*. Wiley-Blackwell, 71(1), pp. 261–277. doi: 10.1002/prot.21715.

Bennion, B. J. and Daggett, V. (2003) 'The molecular basis for the chemical denaturation of proteins by urea.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(9), pp. 5142–7. doi: 10.1073/pnas.0930122100.

Beuckmann, C. T. *et al.* (1999) 'Binding of Biliverdin, Bilirubin, and Thyroid Hormones to Lipocalin-Type Prostaglandin D Synthase', *Biochemistry*. American Chemical Society, 38(25), pp. 8006–8013. doi: 10.1021/BI990261P.

Beynon, R. and Hurst, J. (2003) 'Multiple roles of major urinary proteins in the house mouse, *Mus domesticus*', *Biochemical Society Transactions*. PORTLAND PRESS LTD, THIRD FLOOR, EAGLE HOUSE, 16 PROCTER STREET, LONDON WC1V 6 NX, ENGLAND, 31, pp. 142–146.

Beynon, R. J. *et al.* (2002) 'Polymorphism in major urinary proteins: Molecular heterogeneity in a wild mouse population', *Journal of Chemical Ecology*. SPRINGER, VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS, 28(7), pp. 1429–1446. doi: 10.1023/A:1016252703836.

Beynon, R. J. *et al.* (2008) 'Urinary Lipocalins in Rodenta:is there a Generic Model?', in *Chemical Signals in Vertebrates 11*. New York, NY: Springer New York, pp. 37–49. doi: 10.1007/978-0-387-73945-8_3.

Beynon, R. J. *et al.* (2013) 'The Application of Proteomics to the Discovery and Quantification of

Proteins in Scent Signals', in *Chemical Signals in Vertebrates 12*. New York, NY: Springer New York, pp. 433–447. doi: 10.1007/978-1-4614-5927-9_34.

Beynon, R. J. *et al.* (2014) 'The complexity of protein semiochemistry in mammals.', *Biochemical Society transactions*. Portland Press Limited, 42(4), pp. 837–45. doi: 10.1042/BST20140133.

Beynon, R. J. *et al.* (2015) 'Mass spectrometry for structural analysis and quantification of the Major Urinary Proteins of the house mouse', *International Journal of Mass Spectrometry*, 391, pp. 146–156. doi: 10.1016/j.ijms.2015.07.026.

Beynon, R. J. and Hurst, J. L. (2004) 'Urinary proteins and the modulation of chemical scents in mice and rats.', *Peptides*. ELSEVIER SCIENCE INC, 360 PARK AVE SOUTH, NEW YORK, NY 10010-1710 USA, 25(9), pp. 1553–63. doi: 10.1016/j.peptides.2003.12.025.

Bignetti, E. *et al.* (1985) 'Purification and characterisation of an odorant-binding protein from cow nasal tissue', *European Journal of Biochemistry*, 149, pp. 227–231.

Birke, L. I. A. (1978) 'Scent-marking and the oestrus cycle of the female rat', *Animal Behaviour*, 26, pp. 1165–1166.

Bläker, M. *et al.* (1993) 'Molecular cloning of human von Ebner's gland protein, a member of the lipocalin superfamily highly expressed in lingual salivary glands', *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*. Elsevier, 1172(1–2), pp. 131–137. doi: 10.1016/0167-4781(93)90279-M.

Bligh, E. G. and Dyer, W. J. (1959) 'A rapid method of total lipid extraction and purification', *Canadian Journal of Biochemistry and Physiology*, 37(8), pp. 911–917.

Borowski, Z. (2003) 'Habitat selection and home range size of field voles Microtus agrestis in Słowiński National Park, Poland', *Acta Theriologica*, 48(3), pp. 325–333. doi: 10.1007/BF03194172.

Boschat, C. *et al.* (2002) 'Pheromone detection mediated by a V1r vomeronasal receptor', *Nature Neuroscience*. Nature Publishing Group, 5(12), pp. 1261–1262. doi: 10.1038/nn978.

Boyse, E. A., Beauchamp, G. K. and Yamazaki, K. (1987) 'The genetics of body scent', *Trends in Genetics*. Elsevier Current Trends, 3, pp. 97–102. doi: 10.1016/0168-9525(87)90192-2.

Breed, W. G. (1967) 'Ovulation in the Genus Microtus', *Nature*. Nature Publishing Group, 214(5090), pp. 826–826. doi: 10.1038/214826a0.

Breed, W. G. and Clarke, J. R. (1970) 'Effect of photoperiod on ovarian function in the vole, Microtus agrestis', *Reproduction*, 23(1), pp. 189–192. doi: 10.1530/jrf.0.0230189.

Brennan, P. A. (2009) 'Outstanding issues surrounding vomeronasal mechanisms of pregnancy block and individual recognition in mice', *Behavioural Brain Research*. Elsevier, 200(2), pp. 287–294. doi: 10.1016/J.BBR.2008.10.045.

Briand, Loïc *et al.* (2000) 'Ligand-binding properties and structural characterization of a novel rat odorant-binding protein variant', *European Journal of Biochemistry*. John Wiley & Sons, Ltd (10.1111), 267(10), pp. 3079–3089. doi: 10.1046/j.1432-1033.2000.01340.x.

Briand, Loïc *et al.* (2000) 'Odorant and pheromone binding by aphrodisin, a hamster aphrodisiac protein', *FEBS Letters*, 476(3), pp. 179–185. doi: 10.1016/S0014-5793(00)01719-1.

Briand, L. *et al.* (2004) 'Natural Ligands of Hamster Aphrodisin', *Chemical Senses*. Narnia, 29(5), pp. 425–430. doi: 10.1093/chemse/bjh044.

Briand, L., Trotier, D. and Pernollet, J.-C. (2004) 'Aphrodisin, an aphrodisiac lipocalin secreted in hamster vaginal secretions', *Peptides*, 25(9), pp. 1545–1552. doi: 10.1016/j.peptides.2003.10.026.

Briand, L., Trotier, D. and Pernollet, J. C. (2004) 'Aphrodisin, an aphrodisiac lipocalin secreted in hamster vaginal secretions', *Peptides*, 25(9), pp. 1545–1552. doi: 10.1016/j.peptides.2003.10.026.

Brown, RE; Macdonald, D. (1985) *Social odours in mammals*. Oxford, UK: Clarendon Press.

Brown, K., Innes, J. and Shorten, R. (1993) 'Evidence that possums prey on and scavenge birds' eggs, birds and mammals', *Notornis*, 40(3).

Brown, R. E. (1988) 'Individual odors of rats are discriminable independently of changes in gonadal hormone levels', *Physiology & Behavior*. Elsevier, 43(3), pp. 359–363. doi: 10.1016/0031-9384(88)90199-0.

Brown, R. E. (1995) 'What is the role of the immune system in determining individually distinct body

odours?', *International Journal of Immunopharmacology*. Pergamon, 17(8), pp. 655–661. doi: 10.1016/0192-0561(95)00052-4.

Brownlee, R. G. *et al.* (1969) 'Isolation, Identification and Function of the Chief Component of the Male Tarsal Scent in Black-tailed Deer', *Nature*. Nature Publishing Group, 221(5177), pp. 284–285. doi: 10.1038/221284a0.

Bruce, H. M. (1960) 'A block to pregnancy in the mouse caused by proximity of strange males', *Reproduction*, 1(1), pp. 96–103. doi: 10.1530/jrf.0.0010096.

Brun, V. *et al.* (2007) 'Isotope-labeled protein standards: toward absolute quantitative proteomics.', *Molecular & cellular proteomics : MCP*. American Society for Biochemistry and Molecular Biology, 6(12), pp. 2139–49. doi: 10.1074/mcp.M700163-MCP200.

Brun, V. *et al.* (2009) 'Isotope dilution strategies for absolute quantitative proteomics', *Journal of Proteomics*. Elsevier, 72(5), pp. 740–749. doi: 10.1016/J.JPROT.2009.03.007.

Bujalska, G. (1990) *Social systems and population cycles in voles*. Edited by R. Tamarin et al. International Theriological Congress (5th : 1989 : Rome, Italy): Birkhäuser Verlag.

Burdon, D. W. (1971) 'Immunoglobulins of normal human urine and urethral secretions', *Immunology*, 21(2), pp. 363–368.

Butenandt, A. *et al.* (1959) 'Uber den sexual-lockstoff des seidenspinners Bombyx mori – reindarstellung und konstitution.', *Zeitschrift Fur Naturforschung Part B-Chemie Biochemie Biophysik Biologie Und Verwandten Gebiete*, 14, pp. 283–4.

Butenandt, A., Beckmann, R. and Hecker, E. (1961) 'Über den Sexuallockstoff des Seidenspinners, I. Der biologische Test und die Isolierung des reinen Sexuallockstoffes Bombykol', *Hoppe-Seyler´s Zeitschrift für physiologische Chemie*, 324(1), pp. 71–83.

Cabinet (2016) *Accelerating predator free New Zealand*.

Caley, P. *et al.* (1999) 'Effects of sustained control of brushtail possums on levels of Mycobacterium bovis infection in cattle and brushtail possum populations from Hohotaka, New Zealand', *New Zealand Veterinary Journal*, 47(4), pp. 133–142. doi: 10.1080/00480169.1999.36130.

Cao, Z., Evans, A. R. and Robinson, R. A. S. (2015) 'MS ^{3} -based quantitative proteomics using pulsed-Q dissociation', *Rapid Communications in Mass Spectrometry*. John Wiley & Sons, Ltd, 29(11), pp. 1025–1030. doi: 10.1002/rcm.7192.

Cavaggioni, A. and Mucignat-Caretta, C. (2000) 'Major urinary proteins, α2U-globulins and aphrodisin', *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*. Elsevier, 1482(1–2), pp. 218–228. doi: 10.1016/S0167-4838(00)00149-7.

Chamero, P. *et al.* (2007) 'Identification of protein pheromones that promote aggressive behaviour.', *Nature*. NATURE PUBLISHING GROUP, MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND, 450(7171), pp. 899–902. doi: 10.1038/nature05997.

Chamero, P. *et al.* (2011) 'G protein Gαo is essential for vomeronasal function and aggressive behavior in mice', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(31), p. 12898. doi: 10.1073/PNAS.1107770108.

Chasteen, D. N. (1977) 'Human serotransferrin: structure and function', *Coordination Chemistry Reviews*. Elsevier, 22(1–2), pp. 1–36. doi: 10.1016/S0010-8545(00)80432-4.

Cheetham, S. A. *et al.* (2007) 'The Genetic Basis of Individual-Recognition Signals in the Mouse', *Current Biology*, 17(20), pp. 1771–1777. doi: 10.1016/j.cub.2007.10.007.

Cheetham, S. A. *et al.* (2009) 'Limited variation in the major urinary proteins of laboratory mice', *Physiology & Behavior*. Elsevier, 96(2), pp. 253–261. doi: 10.1016/J.PHYSBEH.2008.10.005.

Chen, V. B. *et al.* (2010) 'MolProbity : all-atom structure validation for macromolecular crystallography', *Acta Crystallographica Section D Biological Crystallography*. International Union of Crystallography, 66(1), pp. 12–21. doi: 10.1107/S0907444909042073.

Chi, H. *et al.* (2010) 'pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra', *Journal of Proteome Research*. American Chemical Society, 9(5), pp. 2713–2724. doi: 10.1021/pr100182k.

Choi, H. *et al.* (2012) 'SAINT-MS1: Protein–Protein Interaction Scoring Using Label-free Intensity Data in Affinity Purification-Mass Spectrometry Experiments', *Journal of Proteome Research*. American

Chemical Society, 11(4), pp. 2619–2624. doi: 10.1021/pr201185r.

Christiansen, E. (1980) 'Urinary marking in wild bank voles, Clethrionomys glareolus in relation to season and sexual status', *Behavioral and Neural Biology*, 28(1), pp. 123–127. doi: 10.1016/S0163-1047(80)93258-6.

Claydon, A. J. *et al.* (2012) 'Protein turnover: Measurement of proteome dynamics by whole animal metabolic labelling with stable isotope labelled amino acids', *Proteomics*. John Wiley & Sons, Ltd, 12(8), pp. 1194–1206. doi: 10.1002/pmic.201100556.

Colby, D. R. and Vandenberg, J. G. (1974) 'Regulatory Effects of Urinary Pheromones on Puberty in the Mouse1', *Biology of Reproduction*. Narnia, 11(3), pp. 268–279. doi: 10.1095/biolreprod11.3.268.

Collet, C., Joseph, R. and Nicholas, K. (1989) 'Molecular cloning and characterization of a novel marsupial milk protein gene', *Biochemical and Biophysical Research Communications*. Academic Press, 164(3), pp. 1380–1383. doi: 10.1016/0006-291X(89)91822-6.

Coombes, H. A., Stockley, P. and Hurst, J. L. (2018) 'Female Chemical Signalling Underlying Reproduction in Mammals', *Journal of chemical ecology*, 44, pp. 851–873. doi: 10.1007/s10886-018-0981-x.

Corner, L. A. L. (2006) 'The role of wild animal populations in the epidemiology of tuberculosis in domestic animals: How to assess the risk', *Veterinary Microbiology*. Elsevier, 112(2–4), pp. 303–312. doi: 10.1016/J.VETMIC.2005.11.015.

Cottrell, J. S. (2011) 'Protein identification using MS/MS data', *Journal of Proteomics*. Elsevier, 74(10), pp. 1842–1851. doi: 10.1016/J.JPROT.2011.05.014.

Couvillon, M. J. *et al.* (2007) 'Nest-mate recognition template of guard honeybees (*Apis mellifera*) is modified by wax comb transfer', *Biology Letters*. The Royal SocietyLondon, 3(3), pp. 228–230. doi: 10.1098/rsbl.2006.0612.

Crawley, M. (1973) 'A live-trapping of Australian brush-tailed possums, Trichosurus vulpecula (Kerr), in the Orongorongo Valley, Wellington, New Zealand', *Australian Journal of Zoology*. CSIRO PUBLISHING, 21(1), p. 75. doi: 10.1071/ZO9730075.

Creasy, D. M. and Cottrell, J. S. (2004) 'Unimod: Protein modifications for mass spectrometry', *Proteomics*. Wiley-Blackwell, 4(6), pp. 1534–1536. doi: 10.1002/pmic.200300744.

Dal Monte, M. *et al.* (1991) 'Purification and characterization of odorant-binding proteins from nasal tissue of rabbit and pig', *Biochem. Physiol*, 99(2), pp. 445–451.

Darwish Marie, A. *et al.* (2001) 'Effect of polymorphisms on ligand binding by mouse major urinary proteins.', *Protein Science*, 10(2), pp. 411–7. doi: 10.1110/ps.31701.

Davis, B. J. (2006) 'Disc electrophoresis- II method and application to human serum proteins', *Annals of the New York Academy of Sciences*. John Wiley & Sons, Ltd (10.1111), 121(2), pp. 404–427. doi: 10.1111/j.1749-6632.1964.tb14213.x.

Desjardins, C., Maruniak, J. A. and Bronson, F. H. (1973) 'Social Rank in House Mice: Differentiation Revealed by Ultraviolet Visualization of Urinary Marking Patterns', *Science*. American Association for the Advancement of Science, pp. 939–941. doi: 10.2307/1737749.

Dey, R. and Roychowdhury, P. (2003) 'Homology model of human corticosteroid binding globulin: a study of its steroid binding ability and a plausible mechanism of steroid hormone release at the site of inflammation', *Journal of Molecular Modeling*. Springer-Verlag, 9(3), pp. 183–189. doi: 10.1007/s00894-003-0130-4.

Dorries, K. M., Adkins-Regan, E. and Halpern, B. P. (1997) 'Sensitivity and Behavioral Responses to the Pheromone Androstenone Are Not Mediated by the Vomeronasal Organ in Domestic Pigs', *Brain, Behavior and Evolution*. Karger Publishers, 49(1), pp. 53–62. doi: 10.1159/000112981.

Drickamer, L. C. (1995) 'Rates of urine excretion by house mouse (*Mus domesticus*): Differences by age, sex, social status, and reproductive condition', *Journal of Chemical Ecology*. Kluwer Academic Publishers-Plenum Publishers, 21(10), pp. 1481–1493. doi: 10.1007/BF02035147.

Drummond, H. *et al.* (2000) 'Competition for Milk in the Domestic Rabbit: Survivors Benefit from Littermate Deaths', *Ethology*. John Wiley & Sons, Ltd (10.1111), 106(6), pp. 511–526. doi: 10.1046/j.1439-0310.2000.00554.x.

Dubey, V. P. *et al.* (2013) 'Copious urinary excretion of a male Syrian hamster (*Mesocricetus auratus*)

salivary gland protein after its endocrine-like release upon β-adrenergic stimulation', *General and Comparative Endocrinology*. Academic Press, 186, pp. 25–32. doi: 10.1016/J.YGCEN.2013.02.016.

Dulac, C. and Axel, R. (1995) 'A novel family of genes encoding putative pheromone receptors in mammals.', *Cell*, 83(2), pp. 195–206.

Eason, C. T. *et al.* (2017) 'Trends in the development of mammalian pest control technology in New Zealand', *New Zealand Journal of Zoology*, 44(4), pp. 267–304. doi: 10.1080/03014223.2017.1337645.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Efford, M. (1998) 'Demographic consequences of sex-biased dispersal in a population of brushtail possums', *Journal of Animal Ecology*. John Wiley & Sons, Ltd, 67(4), pp. 503–517. doi: 10.1046/j.1365-2656.1998.00222.x.

Eisenberg, J. F. and Kleiman, D. G. (1972) 'Olfactory Communication in Mammals', *Annual Review of Ecology and Systematics*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 3(1), pp. 1–32. doi: 10.1146/annurev.es.03.110172.000245.

Elias, J. E. *et al.* (2005) 'Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations', *Nature Methods*, 2(9), pp. 667–675. doi: 10.1038/NMETH785.

Eng, J. K., McCormack, A. L. and Yates, J. R. (1994) 'An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database', *Journal of the American Society for Mass Spectrometry*, 5(11), pp. 976–989. doi: 10.1016/1044-0305(94)80016-2.

Evershed, R. P. *et al.* (1993) 'Application of electrospray ionization mass spectrometry with maximum-entropy analysis to allelic "fingerprinting" of major urinary proteins', *Rapid Communications in Mass Spectrometry*. John Wiley & Sons, Ltd, 7(10), pp. 882–886. doi: 10.1002/rcm.1290071005.

Fabre, A. and Miall, B. (1921) *THe life of Jean Henri Fabre, the antomologist, 1823-1910*. New York: Dodd, Mead and company.

Fan, J. Q. and Lee, Y. C. (1997) 'Detailed studies on substrate structure requirements of glycoamidases A and F.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 272(43), pp. 27058–64. doi: 10.1074/jbc.272.43.27058.

Felicioli, A. *et al.* (1993) 'Multiple types and forms of odorant-binding proteins in the old-world porcupine Hystrix cristata', *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*. Pergamon, 105(3–4), pp. 775–784. doi: 10.1016/0305-0491(93)90119-P.

Felsenstein, J. (1985) 'Confidence limits on phylogenies: an approach using the bootstrap', *Evolution*, 39(4), pp. 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x.

Ferkin, M. H., Lee, D. N. and Leonard, S. T. (2004) 'The Reproductive State of Female Voles Affects their Scent Marking Behavior and the Responses of Male Conspecifics to Such Marks', *Ethology*. John Wiley & Sons, Ltd (10.1111), 110(4), pp. 257–272. doi: 10.1111/j.1439-0310.2004.00961.x.

Ferreira, G. *et al.* (2000) 'Learning of olfactory cues is not necessary for early lamb recognition by the mother', *Physiology & Behavior*. Elsevier, 69(4–5), pp. 405–412. doi: 10.1016/S0031-9384(00)00211-0.

Ferrero, D. M. *et al.* (2011) 'Detection and avoidance of a carnivore odor by prey.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(27), pp. 11235–40. doi: 10.1073/pnas.1103317108.

Ferrero, D. M. *et al.* (2013) 'A juvenile mouse pheromone inhibits sexual behaviour through the vomeronasal system', *Nature*. Nature Publishing Group, 502(7471), pp. 368–371. doi: 10.1038/nature12579.

Finlayson, J. S. *et al.* (1965) 'Major Urinary Protein Complex of Normal Mice: Origin', *Science*. American Association for the Advancement of Science, 149(3687), pp. 981–982. doi: 10.1126/SCIENCE.149.3687.981.

Finlayson, J. S. and Baumann, C. A. (1957) 'Mouse Proteinuria', *American Journal of Physiology-Legacy Content*. American Physiological Society, 192(1), pp. 69–72. doi: 10.1152/ajplegacy.1957.192.1.69.

Fitzgerald, A. E. (1976) 'Diet of the opossum Trichosuvus vulpecula (Kerr) in the Orongorongo Valley, Wellington, New Zealand, in relation to food-plant availability', *New Zealand Journal of Zoology*. Taylor & Francis Group, 3(4), pp. 399–419. doi: 10.1080/03014223.1976.9517928.

Fitzgerald, B. . and Gibb, J. . (2001) 'Introduced mammals in a New Zealand forest: long-term research in the Orongorongo Valley', *Biological Conservation*. Elsevier, 99(1), pp. 97–108. doi: 10.1016/S0006-3207(00)00190-7.

Flower, D. (1996) 'The lipocalin protein family: Structure and function', *Biochemical Journal*. PORTLAND PRESS, 59 PORTLAND PLACE, LONDON, ENGLAND W1N 3AJ, 318, pp. 1–14.

Foster, S. P. (2005) 'Lipid Analysis of the Sex Pheromone Gland of the Moth Heliothis virescens', *Archives of Insect Biochemistry and Physiology*, 59, pp. 80–90. doi: 10.1002/arch.20058.

Frank, A. and Pevzner, P. (2005) 'PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling', *Analytical Chemistry*. American Chemical Society, 77(4), pp. 964–973. doi: 10.1021/AC048788H.

Gaillard, I. *et al.* (2002) 'A single olfactory receptor specifically binds a set of odorant molecules', *European Journal of Neuroscience*. John Wiley & Sons, Ltd (10.1111), 15(3), pp. 409–418. doi: 10.1046/j.0953-816x.2001.01871.x.

Ganfornina, M. D. *et al.* (2000) 'A Phylogenetic Analysis of the Lipocalin Protein Family', *Mol. Biol. Evol*, 17(1), pp. 114–126.

Garibotti, M. *et al.* (1997) 'Three Odorant-binding Proteins from Rabbit Nasal Mucosa', *Chemical Senses*, 22(4), pp. 383–390. doi: 10.1093/chemse/22.4.383.

Garratt, Michael *et al.* (2011) 'Is oxidative stress a physiological cost of reproduction? An experimental test in house mice.', *Proceedings. Biological sciences / The Royal Society*. ROYAL SOC, 6-9 CARLTON HOUSE TERRACE, LONDON SW1Y 5AG, ENGLAND, 278(1708), pp. 1098–106. doi: 10.1098/rspb.2010.1818.

Garratt, M. *et al.* (2011) 'The scent of senescence: Sexual signalling and female preference in house mice', *Journal of Evolutionary Biology*, 24(11), pp. 2398–2409. doi: 10.1111/j.1420-9101.2011.02367.x.

Gasteiger, E. *et al.* (2005) 'Protein Analysis Tools on the ExPASy Server 571 571', in Walker, J. . (ed.) *The Proteomics Protocols Handbook Protein Identification and Analysis Tools on the ExPASy Server*. Totowa, NJ: Humana Press Inc., pp. 571–607. Available at: http://www.expasy.org/tools/. (Accessed: 29 November 2018).

Gerber, S. A. *et al.* (2003) 'Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(12), pp. 6940–6945. doi: 10.1073/pnas.96.12.6591.

GERENA, R. L. *et al.* (2000) 'Stage and Region-Specific Localization of Lipocalin-Type Prostaglandin D Synthase in the Adult Murine Testis and Epididymis', *Journal of Andrology*. John Wiley & Sons, Ltd, 21(6), pp. 848–854. doi: 10.1002/J.1939-4640.2000.TB03415.X.

Getz, W. M. (1991) 'The honey bee as a model kin recognition system', in Hepper, P. G. (ed.) *Kin recognition*. Cambridge: Cambridge University Press, pp. 358–412. doi: 10.1017/CBO9780511525414.015.

Gómez-Baena, G. *et al.* (2014) 'The major urinary protein system in the rat', *Biochemical Society Transactions*, 42(4), pp. 886–892. doi: 10.1042/BST20140083.

Gómez-Baena, G. *et al.* (2019) 'Molecular complexity of the major urinary protein system of the Norway rat, Rattus norvegicus', *Scientific Reports*. Nature Publishing Group, 9(1), p. 10757. doi: 10.1038/s41598-019-46950-x.

Gonzalez-Mariscal, G., Chirino, R. and Hudson, R. (1994) 'Prolactin Stimulates Emission of Nipple Pheromone in Ovariectomized New Zealand White Rabbits', *Biology of Reproduction*, 50(2), pp. 373–376. doi: 10.1095/biolreprod50.2.373.

Gordon, S. *et al.* (1993) *Allergens, IgE, mediators, inflammatory mechanisms Analysis of rat urine proteins and allergens by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and immunoblo~ing*.

Gosling, L. M. and McKay, H. V. (1990) 'Competitor assessment by scent matching: an experimental

test', *Behavioral Ecology and Sociobiology*. Springer-Verlag, 26(6), pp. 415–420. doi: 10.1007/BF00170899.

Gosling, L. M. and Roberts, S. C. (2001) 'Scent-marking by male mammals: Cheat-proof signals to competitors and mates', *Advances in the Study of Behavior*, 30, pp. 169–217. doi: 10.1016/S0065-3454(01)80007-3.

Gouveia, K. and Hurst, J. L. (2017) 'Optimising reliability of mouse performance in behavioural testing: the major role of non-aversive handling', *Scientific Reports*. Nature Publishing Group, 7(1), p. 44999. doi: 10.1038/srep44999.

Green, W. and Rohan, M. (2011) 'Opposition to aerial 1080 poisoning for control of invasive mammals in New Zealand: risk perceptions and agency responses', *Journal of the Royal Society of New Zealand*, 42(3), pp. 185–213. doi: 10.1080/03036758.2011.556130.

Greene, T. C. *et al.* (2013) 'Monitoring selected forest bird species through aerial application of 1080 baits, Waitutu, New Zealand', *New Zealand Journal of Ecology*, 37(1), pp. 41–50.

Gregoire, C. *et al.* (1996) 'cDNA Cloning and Sequencing Reveal the Major Horse Allergen Equ c1 to Be a Glycoprotein Member of the Lipocalin Superfamily', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 271(51), pp. 32951–32959. doi: 10.1074/jbc.271.51.32951.

de Groot, H. *et al.* (1991) 'Affinity purification of a major and a minor allergen from dog extract: serologic activity of affinity-purified Can f I and of Can f I-depleted extract.', *The Journal of allergy and clinical immunology*, 87(6), pp. 1056–65.

Haga, S. *et al.* (2010) 'The male mouse pheromone ESP1 enhances female sexual receptive behaviour through a specific vomeronasal receptor', *Nature*. Nature Publishing Group, 466(7302), pp. 118–122. doi: 10.1038/nature09142.

Hagemeyer, P. *et al.* (2011) 'Searching for Major Urinary Proteins (MUPs) as Chemosignals in Urine of Subterranean Rodents', *Journal of Chemical Ecology*. Springer-Verlag, 37(7), pp. 687–694. doi: 10.1007/s10886-011-9971-y.

Hagenäs, L. *et al.* (1975) 'Sertoli cell origin of testicular androgen-binding protein (ABP)', *Molecular and Cellular Endocrinology*, 2(5), pp. 339–350. doi: 10.1016/0303-7207(75)90021-0.

Hagiwara, K. *et al.* (2003) 'Mouse SWAM1 and SWAM2 Are Antibacterial Proteins Composed of a Single Whey Acidic Protein Motif', *The Journal of Immunology*. American Association of Immunologists, 170(4), pp. 1973–1979. doi: 10.4049/jimmunol.170.4.1973.

Halle, S. and Lehmann, U. (1992) 'Cycle-Correlated Changes in the Activity Behaviour of Field Voles, *Microtus agrestis*', *Oikos*, 64(3), p. 489. doi: 10.2307/3545166.

Hammond, G. L. (1990) 'Molecular Properties of Corticosteroid Binding Globulin and the Sex-Steroid Binding Proteins', *Endocrine Reviews*. Oxford University Press, 11(1), pp. 65–79. doi: 10.1210/edrv-11-1-65.

Han, X. *et al.* (2011) 'PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications', *Journal of Proteome Research*. American Chemical Society, 10(7), pp. 2930–2936. doi: 10.1021/pr200153k.

Han, Y., Ma, B. and Zhang, K. (2005) 'SPIDER: Software for protein identification from sequence tags with de novo sequencing error', *Journal of Bioinformatics and Computational Biology*. Imperial College Press, 03(03), pp. 697–716. doi: 10.1142/S0219720005001247.

Hanski, I., Hansson, L. and Henttonen, H. (1991) 'Specialist Predators, Generalist Predators, and the Microtine Rodent Cycle', *The Journal of Animal Ecology*, 60(1), p. 353. doi: 10.2307/5465.

Hansson, L. (1986) 'Geographic differences in the sociability of voles in relation to cyclicity', *Animal Behaviour*. Academic Press, 34(4), pp. 1215–1221. doi: 10.1016/S0003-3472(86)80181-6.

Hansson, L. and Henttonen, H. (1985) 'Gradients in density variations of small rodents: the importance of latitude and snow cover', *Oecologia*. Springer-Verlag, 67(3), pp. 394–402. doi: 10.1007/BF00384946.

Hara-Kuge, S. *et al.* (2002) 'Involvement of VIP36 in intracellular transport and secretion of glycoproteins in polarized Madin-Darby canine kidney (MDCK) cells.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 277(18), pp. 16332–9. doi:

10.1074/jbc.M112188200.

Hastie, N. D. and Held, W. A. (1978) *Analysis of mRNA populations by cDNA*mRNA hybrid-mediated inhibition of cell-free protein synthesis (globin mRNA/hybridization/translation/abundant liver mRNAs)*.

Hattori, T. *et al.* (2016) 'Self-Exposure to the Male Pheromone ESP1 Enhances Male Aggressiveness in Mice', *Current Biology*, 26, pp. 1229–1234. doi: 10.1016/j.cub.2016.03.029.

Hattori, T. *et al.* (2017) 'Exocrine Gland-Secreting Peptide 1 Is a Key Chemosensory Signal Responsible for the Bruce Effect in Mice.', *Current Biology*. Elsevier, 27(20), pp. 3197-3201.e3. doi: 10.1016/j.cub.2017.09.013.

Havilio, M. and Wool, A. (2007) 'Large-Scale Unrestricted Identification of Post-Translation Modifications Using Tandem Mass Spectrometry', *Analytical Chemistry*. American Chemical Society, 79(4), pp. 1362–1368. doi: 10.1021/ac061515x.

He, B. *et al.* (2019) 'Label-free absolute protein quantification with data-independent acquisition', *Journal of Proteomics*. Elsevier, 200, pp. 51–59. doi: 10.1016/J.JPROT.2019.03.005.

Held, W. A. and Gallagher, J. F. (1985) 'Rat α2u-globulin mRNA expression in the preputial gland', *Biochemical Genetics*. Kluwer Academic Publishers-Plenum Publishers, 23(3–4), pp. 281–290. doi: 10.1007/BF00504325.

Henzel, W. J. *et al.* (1988) 'The primary structure of aphrodisin.', *The Journal of biological chemistry*, 263(32), pp. 16682–7.

Herman, J. S. *et al.* (2019) 'Genetic variation in field voles (*Microtus agrestis*) from the British Isles: selective sweeps or population bottlenecks?', *Biological Journal of the Linnean Society*. Narnia, 126(4), pp. 852–865. doi: 10.1093/biolinnean/bly213.

Herrada, G. and Dulac, C. (1997) 'A novel family of putative pheromone receptors in mammals with a topographically organized and sexually dimorphic distribution.', *Cell*. Elsevier, 90(4), pp. 763–73. doi: 10.1016/S0092-8674(00)80536-X.

Hershko, A., Ciechanover, A. and Varshavsky, A. (2000) 'The ubiquitin system', *Nature Medicine*. Nature Publishing Group, 6(10), pp. 1073–1081. doi: 10.1038/80384.

Heydel, J.-M. *et al.* (2013) 'Odorant-Binding Proteins and Xenobiotic Metabolizing Enzymes: Implications in Olfactory Perireceptor Events', *The Anatomical Record*. John Wiley & Sons, Ltd, 296(9), pp. 1333–1345. doi: 10.1002/ar.22735.

Ho, C. S. *et al.* (2003) 'Electrospray ionisation mass spectrometry: principles and clinical applications.', *The Clinical biochemist. Reviews*. The Australian Association of Clinical Biochemists, 24(1), pp. 3–12.

Hoffmann, A. *et al.* (1996) 'Developmental expression of murine β-trace in embryos and adult animals suggests a function in maturation and maintenance of blood-tissue barriers', *Developmental Dynamics*. John Wiley & Sons, Ltd, 207(3), pp. 332–343. doi: 10.1002/(SICI)1097-0177(199611)207:3<332::AID-AJA10>3.0.CO;2-6.

Holmes, W. G. (1986) 'Identification of paternal half-siblings by captive Belding's ground squirrels', *Animal Behaviour*. Academic Press, 34(2), pp. 321–327. doi: 10.1016/S0003-3472(86)80099-9.

Holmes, W. G. and Sherman, P. W. (1982) *The Ontogeny of Kin Recognition in Two Species of Ground Squirrels 1*.

Hu, H. *et al.* (2016) 'A review of methods for interpretation of glycopeptide tandem mass spectral data', *Glycoconjugate Journal*. Springer US, 33(3), pp. 285–296. doi: 10.1007/s10719-015-9633-3.

Hudson, R. and Distel, H. (1983) 'Nipple Location By Newborn Rabbits: Behavioural Evidence for Pheromonal Guidance', *Behaviour*, 85(3–4). doi: https://doi.org/10.1163/156853983X00255.

Hudson, R., González-Mariscal, G. and Beyer, C. (1990) 'Chin marking behavior, sexual receptivity, and pheromone emission in steroid-treated, ovariectomized rabbits', *Hormones and Behavior*. Academic Press, 24(1), pp. 1–13. doi: 10.1016/0018-506X(90)90022-P.

Hudson, R. and Vodermayer, T. (1992) 'Spontaneous and odour-induced chin marking in domestic female rabbits', *Animal Behaviour*, 43, pp. 329–336.

Hurst, H. C. and Parker, M. G. (1983) 'Rat prostatic steroid binding protein: DNA sequence and transcript maps of the two C3 genes.', *The EMBO journal*. European Molecular Biology Organization,

2(5), pp. 769–74.

Hurst, J. and Beynon, R. (2004) 'Scent wars: the chemobiology of competitive signalling in mice', *BIOESSAYS*. WILEY-BLACKWELL, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 26(12), pp. 1288–1298. doi: 10.1002/bles.20147.

Hurst, J. L. (1990) 'Urine marking in populations of wild house mice Mus domesticus rutty. I. Communication between males', *Animal Behaviour*. Academic Press, 40(2), pp. 209–222. doi: 10.1016/S0003-3472(05)80916-9.

Hurst, J. L. (1993) 'The priming effects of urine substrate marks on interactions between male house mice, Mus musculus domesticus Schwarz & Schwarz', *Animal Behaviour*, 45(1), pp. 55–81. doi: 10.1006/anbe.1993.1007.

Hurst, J. L. *et al.* (1998) 'Proteins in urine scent marks of male house mice extend the longevity of olfactory signals', *Animal Behaviour*. ACADEMIC PRESS LTD, 24-28 OVAL RD, LONDON NW1 7DX, ENGLAND, 55(5), pp. 1289–1297. doi: 10.1006/anbe.1997.0650.

Hurst, J. L. *et al.* (2001) 'Individual recognition in mice mediated by major urinary proteins.', *Nature*. NATURE PUBLISHING GROUP, MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND, 414(6864), pp. 631–4. doi: 10.1038/414631a.

Hurst, J. L. *et al.* (2005) 'MHC odours are not required or sufficient for recognition of individual scent owners', *Proceedings of the Royal Society B: Biological Sciences*, 272(1564), pp. 715–724. doi: 10.1098/rspb.2004.3004.

Hurst, J. L. *et al.* (2017) 'Molecular heterogeneity in major urinary proteins of Mus musculus subspecies: potential candidates involved in speciation', *Scientific Reports*, 7(1), p. 44992. doi: 10.1038/srep44992.

Hurst, J. L. and Beynon, R. J. (2013) 'Rodent Urinary Proteins: Genetic Identity Signals and Pheromones', in East, M. and Dehnhard, M. (eds) *Chemical Signals in Vertebrates 12*. New York, NY: Springer, pp. 117–133. doi: 10.1007/978-1-4614-5927-9_9.

Hwang, J. M. *et al.* (1997) 'The microevolution of mouse salivary androgen-binding protein (ABP) paralleled subspeciation of Mus musculus', *Journal of Heredity*. Narnia, 88(2), pp. 93–97. doi: 10.1093/oxfordjournals.jhered.a023083.

Ilmonen, P. *et al.* (2007) 'Major histocompatibility complex heterozygosity reduces fitness in experimentally infected mice.', *Genetics*. Genetics, 176(4), pp. 2501–8. doi: 10.1534/genetics.107.074815.

Ilmonen, P. *et al.* (2009) 'Females prefer the scent of outbred males: good-genes-as-heterozygosity?', *BMC Evolutionary Biology*, 9(104). doi: 10.1186/1471-2148-9-104.

Innes, J. G. *et al.* (2005) *Kokako population studies at Rotoehu Forest and on Little Barrier*, *Science for Conservation*. Wellington, New Zealand.

Ishii, T., Hirota, J. and Mombaerts, P. (2003) 'Combinatorial coexpression of neural and immune multigene families in mouse vomeronasal sensory neurons.', *Current Biology*. Elsevier, 13(5), pp. 394–400. doi: 10.1016/S0960-9822(03)00092-7.

Ishii, T. and Mombaerts, P. (2008) 'Expression of nonclassical class I major histocompatibility genes defines a tripartite organization of the mouse vomeronasal system.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*. Society for Neuroscience, 28(10), pp. 2332–41. doi: 10.1523/JNEUROSCI.4807-07.2008.

Isogai, Y. *et al.* (2011) 'Molecular organization of vomeronasal chemoreception', *Nature*. Nature Publishing Group, 478(7368), pp. 241–245. doi: 10.1038/nature10437.

IUCN (2018) 'The IUCN Red List of Threatened Species 2018', p. Version 2018-2. Available at: https://www.iucnredlist.org (Accessed: 28 November 2018).

Jemiolo, B. *et al.* (1985) 'Behavioural and endocrine responses of female mice to synthetic analogues of volatile compounds in male urine', *Animal Behaviour*. Academic Press, 33(4), pp. 1114–1118. doi: 10.1016/S0003-3472(85)80170-6.

Jemiolo, B. and Novotny, M. (1994) 'Inhibition of sexual maturation in juvenile female and male mice by a chemosignal of female origin', *Physiology & Behavior*. Elsevier, 55(3), pp. 519–522. doi: 10.1016/0031-9384(94)90110-4.

Jemiolo, B., Xie, T.-M. and Novotny, M. (1991) 'Socio-sexual olfactory preference in female mice: Attractiveness of synthetic chemosignals', *Physiology & Behavior*. Elsevier, 50(6), pp. 1119–1122. doi: 10.1016/0031-9384(91)90570-E.

Ji, W., White, P. C. L. and Clout, M. N. (2005) 'Contact rates between possums revealed by proximity data loggers', *Journal of Applied Ecology*. Wiley/Blackwell (10.1111), 42(3), pp. 595–604. doi: 10.1111/j.1365-2664.2005.01026.x.

Johnson, M. A. *et al.* (2000) 'Isolation and characterization of mouse probasin: An androgen-regulated protein specifically expressed in the differentiated prostate', *The Prostate*. John Wiley & Sons, Ltd, 43(4), pp. 255–262. doi: 10.1002/1097-0045(20000601)43:4<255::AID-PROS4>3.0.CO;2-M.

Johnson, R. P. (1973) 'Scent marking in mammals', *Animal Behaviour*, 21(3), pp. 521–535. doi: 10.1016/S0003-3472(73)80012-0.

Johnson, R. P. (1975) 'Scent Marking With Urine in Two Races of the Bank Vole (*Clethrionomys glareolus*)', *Behaviour*. E J BRILL, PO BOX 9000, 2300 PA LEIDEN, NETHERLANDS, 55(1), pp. 81–93. doi: 10.1163/156853975X00425.

Johnston, R. E. (1977) 'The causation of two scent-marking behaviour patterns in female hamsters (Mesocricetus auratus)', *Animal Behaviour*. Academic Press, 25, pp. 317–327. doi: 10.1016/0003-3472(77)90007-0.

Jolly, S. E., Scobie, S. and Coleman, M. C. (1995) 'Breeding capacity of female brushtail possums Trichosurus vulpecula in captivity', *New Zealand Journal of Zoology*, 22, pp. 325–330. doi: 10.1080/03014223.1995.9518048.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) 'The rapid generation of mutation data matrices from protein sequences.', *Computer applications in the biosciences : CABIOS*, 8(3), pp. 275–82.

Kapusta, J. and Sales, G. (2009) 'Male–female interactions and ultrasonic vocalization in three sympatric species of voles during conspecific and heterospecific encounters', *Behaviour*, 146(7), pp. 939–962. doi: https://doi.org/10.1163/156853908X396818.

Karn, R. C. and Laukaitis, C. M. (2009) 'The Mechanism of Expansion and the Volatility it created in Three Pheromone Gene Clusters in the Mouse (*Mus musculus*) Genome', *Genome Biology and Evolution*, 1, pp. 494–503. doi: 10.1093/gbe/evp049.

Kasper, S. and Matusik, R. J. (2000) 'Rat probasin: structure and function of an outlier lipocalin', *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*. Elsevier, 1482(1–2), pp. 249–258. doi: 10.1016/S0167-4838(00)00170-9.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability.', *Molecular biology and evolution*. Oxford University Press, 30(4), pp. 772–80. doi: 10.1093/molbev/mst010.

Kaur, A. W. *et al.* (2014) 'Murine pheromone proteins constitute a context-dependent combinatorial code governing multiple social behaviors.', *Cell*, 157(3), pp. 676–88. doi: 10.1016/j.cell.2014.02.025.

Kavaliers, M. and Colwellt, D. D. (1995) 'Odours of parasitized males induce aversive responses in female mice', *Animal Behaviour*, 50, pp. 1161–169.

Kawabata, T. (2010) 'Detection of multiscale pockets on protein surfaces using mathematical morphology', *Proteins: Structure, Function, and Bioinformatics*. John Wiley & Sons, Ltd, 78(5), pp. 1195–1211. doi: 10.1002/prot.22639.

Keller, A. *et al.* (2007) 'Genetic variation in a human odorant receptor alters odour perception', *Nature*. Nature Publishing Group, 449(7161), pp. 468–472. doi: 10.1038/nature06162.

Kelley, L. A. *et al.* (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis.', *Nature protocols*. Europe PMC Funders, 10(6), pp. 845–58. doi: 10.1038/nprot.2015.053.

Kemp, J. R. *et al.* (2018) 'Effects of the aerial application of 1080 to control pest mammals on kea reproductive success', *New Zealand Journal of Ecology*, 42(2), pp. 158–168. doi: 10.20417/nzjecol.42.28.

Kerr, R. (1792) *The animal kingdom, or zoological system, of the celebrated Sir Charles Linnæus.* Edinburgh : Printed for A. Strahan, and T. Cadell, London, and W. Creech, Edinburgh,. doi: 10.5962/bhl.title.57940.

Kidd, B. A., Baker, D. and Thomas, W. E. (2009) 'Computation of Conformational Coupling in

Allosteric Proteins', *PLoS Computational Biology*. Edited by A. Horovitz. Public Library of Science, 5(8), p. e1000484. doi: 10.1371/journal.pcbi.1000484.

Kimoto, H. *et al.* (2005) 'Sex-specific peptides from exocrine glands stimulate mouse vomeronasal sensory neurons.', *Nature*. NATURE PUBLISHING GROUP, MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND, 437(7060), pp. 898–901. doi: 10.1038/nature04033.

Kimoto, H. *et al.* (2007) 'Sex- and strain-specific expression and vomeronasal activity of mouse ESP family peptides.', *Current biology : CB*. Elsevier, 17(21), pp. 1879–84. doi: 10.1016/j.cub.2007.09.042.

Knopf, J. L., Gallagher, J. F. and Held, W. A. (1983) 'Differential, Multihormonal Regulation of the Mouse Major Urinary Protein Gene Family in the Liver Downloaded from', *Molecular & Cellular Proteomics*, 3(12), pp. 2232–2240.

Kobayakawa, K. *et al.* (2007) 'Innate versus learned odour processing in the mouse olfactory bulb', *Nature*. Nature Publishing Group, 450(7169), pp. 503–508. doi: 10.1038/nature06281.

Kong, A. T. *et al.* (2017) 'MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics', *Nature Methods*, 14(5), pp. 513–520. doi: 10.1038/nmeth.4256.

Koyama, S. (2004) 'Primer effects by conspecific odors in house mice: a new perspective in the study of primer effects on reproductive activities'. doi: 10.1016/j.yhbeh.2004.03.002.

Krissinel, E. (2012) 'Enhanced fold recognition using efficient short fragment clustering.', *Journal of molecular biochemistry*. Europe PMC Funders, 1(2), pp. 76–85.

Kruczek, M. (1997) 'Male rank and female choice in the bank vole, *Clethrionomys glareolus*', *Behavioural Processes*. ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS, 40(2), pp. 171–176. doi: 10.1016/S0376-6357(97)00785-7.

Kruczek, M. and Gołas, A. (2003) 'Behavioural development of conspecific odour preferences in bank voles, Clethrionomys glareolus', *Behavioural Processes*. ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS, 64(1), pp. 31–39. doi: 10.1016/S0376-6357(03)00107-4.

Kruczek, M. and Marchlewskakoj, A. (1985) 'Androgen-dependent proteins in the urine of bank voles (*Clethrionomys-glareolus*)', *Journal of Reproduction and Fertility*. J REPROD FERTIL INC, 22 NEWMARKET RD, CAMBRIDGE, ENGLAND CB5 8DT, 75(1), pp. 189–192.

Kruuk, H., Gorman, M. and Leitch, A. (1984) 'Scent-marking with the subcaudal gland by the European badger, *Meles meles L.*', *Animal Behaviour*. Academic Press, 32(3), pp. 899–907. doi: 10.1016/S0003-3472(84)80168-2.

Laemmli, U. K. (1970) 'Cleavage of structural proteins during the assembly of the head of bacteriophage T4.', *Nature*, 227(5259), pp. 680–5.

Lai, S.-C., Vasilieva, N. Y. and Johnston, R. E. (1996) 'Odors Providing Sexual Information in Djungarian Hamsters: Evidence for an Across-Odor Code', *Hormones and Behavior*, 30, pp. 26–36.

Lambin, X., Bretagnolle, V. and Yoccoz, N. G. (2006) 'Vole population cycles in northern and southern Europe: Is there a need for different explanations for single pattern?', *Journal of Animal Ecology*. John Wiley & Sons, Ltd (10.1111), 75(2), pp. 340–349. doi: 10.1111/j.1365-2656.2006.01051.x.

Laskowski, R. A. *et al.* (1993) 'PROCHECK: a program to check the stereochemical quality of protein structures', *Journal of Applied Crystallography*. International Union of Crystallography, 26(2), pp. 283–291. doi: 10.1107/S0021889892009944.

Laskowski, R. A., Watson, J. D. and Thornton, J. M. (2005) 'ProFunc: a server for predicting protein function from 3D structure.', *Nucleic acids research*. Oxford University Press, 33(Web Server issue), pp. W89-93. doi: 10.1093/nar/gki414.

Laukaitis, C. M. *et al.* (2005) 'Diverse spatial, temporal, and sexual expression of recently duplicated androgen-binding protein genes in Mus musculus', *BMC Evolutionary Biology*. BioMed Central, 5(1), p. 40. doi: 10.1186/1471-2148-5-40.

Laukaitis, C. M., Critser, E. S. and Karn, R. C. (1997) 'Salivary androgen-binding protein (ABP) mediates sexual isolation in Mus musculus', *Evolution*. John Wiley & Sons, Ltd (10.1111), 51(6), pp. 2000–2005. doi: 10.1111/j.1558-5646.1997.tb05121.x.

Lazar, J. *et al.* (2002) 'Molecular and Functional Characterization of an Odorant Binding Protein of the Asian Elephant, Elephas maximus: Implications for the Role of Lipocalins in Mammalian Olfaction', *Biochemistry*. American Chemical Society, 41(39), pp. 11786–11794. doi: 10.1021/BI0256734.

Lebedev, A. T. *et al.* (2014) 'Discrimination of Leucine and Isoleucine in Peptides Sequencing with Orbitrap Fusion Mass Spectrometer', *Analytical Chemistry*. American Chemical Society, 86(14), pp. 7017–7022. doi: 10.1021/ac501200h.

Van der Lee, S. and Boot, L. M. (1956) 'Spontaneous pseudopregnancy in mice. II.', *Acta physiologica et pharmacologica Neerlandica*, 5(2), pp. 213–5.

Lee, V. (2015) *The Application of Protein Mass Spectrometry to the Understanding of Behaviour in Mus Species*. University of Liverpool.

Leinders-Zufall, T. *et al.* (2000) 'Ultrasensitive pheromone detection by mammalian vomeronasal neurons', *Nature*. Nature Publishing Group, 405(6788), pp. 792–796. doi: 10.1038/35015572.

Leinders-Zufall, T. *et al.* (2004) 'MHC Class I Peptides as Chemosensory Signals in the Vomeronasal Organ', *Science*. American Association for the Advancement of Science, 306(5698), pp. 1033–1037. doi: 10.1126/SCIENCE.1102818.

Leinders-Zufall, T. *et al.* (2009) 'Structural requirements for the activation of vomeronasal sensory neurons by MHC peptides', *Nature Neuroscience*. Nature Publishing Group, 12(12), pp. 1551–1558. doi: 10.1038/nn.2452.

Lévy, F. *et al.* (1995) 'Involvement of the main but not the accessory olfactory system in maternal behavior of primiparous and multiparous ewes', *Physiology & Behavior*. Elsevier, 57(1), pp. 97–104. doi: 10.1016/0031-9384(94)00200-O.

Lévy, F. and Keller, M. (2008) 'Neurobiology of Maternal Behavior in Sheep', in Brockmann, H. J. et al. (eds) *Advances in the Study of Behavior*. Academic Press, pp. 399–437. doi: 10.1016/S0065-3454(08)00008-9.

Lévy, F. and Keller, M. (2009) 'Olfactory mediation of maternal behavior in selected mammalian species', *Behavioural Brain Research*. Elsevier, 200(2), pp. 336–345. doi: 10.1016/J.BBR.2008.12.017.

Li, Q. *et al.* (2013) 'Synchronous Evolution of an Odor Biosynthesis Pathway and Behavioral Response', *Current Biology*. Elsevier, 23(1), pp. 11–20. doi: 10.1016/j.cub.2012.10.047.

Li, Z. *et al.* (2012) 'Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos', *Journal of Proteome Research*. American Chemical Society, 11(3), pp. 1582–1590. doi: 10.1021/pr200748h.

Liberles, S. D. (2014) 'Mammalian Pheromones', *Annual Review of Physiology*. Annual Reviews, 76(1), pp. 151–175. doi: 10.1146/annurev-physiol-021113-170334.

Liberles, S. D. and Buck, L. B. (2006) 'A second class of chemosensory receptors in the olfactory epithelium', *Nature*. Nature Publishing Group, 442(7103), pp. 645–650. doi: 10.1038/nature05066.

Liman, E. R. and Innan, H. (2003) 'Relaxed selective pressure on an essential component of pheromone transduction in primate evolution.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(6), pp. 3328–32. doi: 10.1073/pnas.0636123100.

Lin, D. Y. *et al.* (2005) 'Encoding social signals in the mouse main olfactory bulb', *Nature*. Nature Publishing Group, 434(7032), pp. 470–477. doi: 10.1038/nature03414.

Lin, H.-Y. and Muller, Y. A. (2010) 'Molecular and structural basis of steroid hormone binding and release from corticosteroid-binding globulin', *Molecular and Cellular Endocrinology*. Elsevier, 316(1), pp. 3–12. doi: 10.1016/J.MCE.2009.06.015.

Linnaeus, C. (1761) *Fauna svecica : sistens animalia sveciae regni: mammalia, aves, amphibia, pisces, insecta, vermes, distributa per classes &; ordines, genera &; species, cum differentiis specierum, synonymis auctorum, nominibus incolarum, locis natalium, description*. Stockholmiae : Sumtu & Literis Direct. Laurentii Salvii,.

Lisk, R. D. and Nachtigall, M. J. (1988) 'Estrogen regulation of agonistic and proceptive responses in the golden hamster', *Hormones and Behavior*. Academic Press, 22(1), pp. 35–48. doi: 10.1016/0018-506X(88)90029-3.

Lobel, D. *et al.* (2001) 'Identification of a Third Rat Odorant-binding Protein (OBP3)', *Chemical Senses*. Oxford University Press, 26(6), pp. 673–680. doi: 10.1093/chemse/26.6.673.

Loconto, J. *et al.* (2003) 'Functional Expression of Murine V2R Pheromone Receptors Involves Selective Association with the M10 and M1 Families of MHC Class Ib Molecules', *Cell*. Elsevier,

112(5), pp. 607–618. doi: 10.1016/S0092-8674(03)00153-3.

Loebel, D. *et al.* (2000) 'Cloning, post-translational modifications, heterologous expression and ligand-binding of boar salivary lipocalin.', *The Biochemical journal*. Portland Press Ltd, 350 Pt 2(Pt 2), pp. 369–79. doi: 10.1042/0264-6021:3500369.

Logan, D. W., Marton, T. F. and Stowers, L. (2008) 'Species specificity in major urinary proteins by parallel evolution.', *PloS one*. PUBLIC LIBRARY SCIENCE, 185 BERRY ST, STE 1300, SAN FRANCISCO, CA 94107 USA, 3(9), p. e3280. doi: 10.1371/journal.pone.0003280.

Lopes, P. C. and König, B. (2016) 'Choosing a healthy mate: sexually attractive traits as reliable indicators of current disease status in house mice', *Animal Behaviour*, 111, pp. 119–126. doi: 10.1016/j.anbehav.2015.10.011.

Loughran, M. F. E. (2006) 'Social organisation of female field voles *Microtus agrestis* in a population in Southern England', *Acta Theriologica*. Springer-Verlag, 51(3), pp. 233–242. doi: 10.1007/BF03192675.

Loxley, G. M. *et al.* (2017) 'Glareosin : a novel sexually dimorphic urinary lipocalin in the bank vole , *Myodes glareolus*', *Open Biology*, 7(9), p. 170135. doi: 10.1098/rsob.170135.

Ma, B. *et al.* (2003) 'PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry', *Rapid Communications in Mass Spectrometry*. John Wiley & Sons, Ltd, 17(20), pp. 2337–2342. doi: 10.1002/rcm.1196.

Maccioni, M., Riera, C. . and Rivero, V. . (2001) 'Identification of rat prostatic steroid binding protein (PSBP) as an immunosuppressive factor', *Journal of Reproductive Immunology*. Elsevier, 50(2), pp. 133–149. doi: 10.1016/S0165-0378(01)00060-2.

Macrides, F., Bartke, A. and Dalterio, S. (1975) 'Strange females increase plasma testosterone levels in male mice.', *Science*. American Association for the Advancement of Science, 189(4208), pp. 1104–6. doi: 10.1126/science.1162363.

Maley, F. *et al.* (1989) 'Characterization of glycoproteins and their associated oligosaccharides through the use of endoglycosidases', *Analytical Biochemistry*. Academic Press, 180(2), pp. 195–204. doi: 10.1016/0003-2697(89)90115-2.

Mandelli, M.-J. and Sales, G. (2004) 'Ultrasonic vocalizations of infant short-tailed field voles, *Microtus agrestis*', *Journal of Mammalogy*. Narnia, 85(2), pp. 282–289. doi: 10.1644/1545-1542(2004)085<0282:UVOISF>2.0.CO;2.

Mann, M. and Wilm, M. (1994) 'Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags', *Analytical Chemistry*. American Chemical Society, 66(24), pp. 4390–4399. doi: 10.1021/ac00096a002.

Marchese, S. *et al.* (1998) 'Lipocalins of boar salivary glands binding odours and pheromones', *European Journal of Biochemistry*, 252(3), pp. 563–568. doi: 10.1046/j.1432-1327.1998.2520563.x.

Martin, H. (2011) 'Laboratory measurement of urine albumin and urine total protein in screening for proteinuria in chronic kidney disease.', *The Clinical biochemist. Reviews*. The Australian Association of Clinical Biochemists, 32(2), pp. 97–102.

Martínez-Marcos, A. and Halpern, M. (2009) 'Evolution of Olfactory and Vomeronasal Systems', in *Encyclopedia of Neuroscience*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1264–1269. doi: 10.1007/978-3-540-29678-2_3135.

Martini, S. *et al.* (2008) 'Co-Expression of Putative Pheromone Receptors in the Sensory Neurons of the Vomeronasal Organ', *Journal of Neuroscience*. Society for Neuroscience, 21(3), pp. 843–848. doi: 20026321.

Massey, A. *et al.* (1980) 'Puberty delay by a urinary cue from female house mice in feral populations', *Science*. American Association for the Advancement of Science, 209(4458), pp. 821–822. doi: 10.1126/science.7190728.

Mastrogiacomo, R. *et al.* (2014) 'An Odorant-Binding Protein Is Abundantly Expressed in the Nose and in the Seminal Fluid of the Rabbit', *PLoS ONE*. Edited by S. D'Auria, 9(11), p. e111932. doi: 10.1371/journal.pone.0111932.

Mateo, J. M. (2009) 'The causal role of odours in the development of recognition templates and social preferences', *Animal Behaviour*. Academic Press, 77(1), pp. 115–121. doi:

10.1016/J.ANBEHAV.2008.09.015.

Matuo, Y. *et al.* (1982) 'Isolation and characterization of androgen-dependent non-histone chromosomal protein from dorsolateral prostate of rats', *Biochemical and Biophysical Research Communications*. Academic Press, 109(2), pp. 334–340. doi: 10.1016/0006-291X(82)91725-9.

Matuo, Y. *et al.* (1984) 'Changes of an androgen-dependent nuclear protein during functional differentiation and by dedifferentiation of the dorsolateral prostate of rats', *Biochemical and Biophysical Research Communications*. Academic Press, 118(2), pp. 467–473. doi: 10.1016/0006-291X(84)91326-3.

Mazurkiewicz, M. (1981) *Spatial-organization of a bank vole population in years of small or large numbers*. POLISH ACAD SCIENCES, MAMMAL RESEARCH INST, 17-230 BIALOWIEZA, POLAND: Państwowe Wydawn. Naukowe.

McKenzie, H. A., Muller, V. J. and Treacy, G. B. (1983) '"Whey" proteins of milk of the red (Macropus rufus) and eastern grey (Macropus giganteus) kangaroo', *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*. Pergamon, 74(2), pp. 259–271. doi: 10.1016/0305-0491(83)90010-X.

McLean, L. *et al.* (2007) 'Characterization of Cauxin in the Urine of Domestic and Big Cats', *Journal of Chemical Ecology*. Springer-Verlag, 33(10), pp. 1997–2009. doi: 10.1007/s10886-007-9354-6.

McLean, S. (2014) 'Scent glands of the common brushtail possum (*Trichosurus vulpecula*)', *New Zealand Journal of Zoology*, 41(3), pp. 193–202. doi: 10.1080/03014223.2014.899506.

McLean, S. *et al.* (2015) 'Triacylglycerol Estolides, a New Class of Mammalian Lipids, in the Paracloacal Gland of the Brushtail Possum (*Trichosurus vulpecula*)', *Lipids*. Springer Berlin Heidelberg, 50(6), pp. 591–604. doi: 10.1007/s11745-015-4025-9.

McLean, S., Davies, N. W. and Wiggins, N. L. (2012) 'Scent Chemicals of the Brushtail Possum, Trichosurus vulpecula', *Journal of Chemical Ecology*. Springer-Verlag, 38(10), pp. 1318–1339. doi: 10.1007/s10886-012-0188-5.

Mechref, Y. *et al.* (2000) 'Glycosylated major urinary protein of the house mouse: characterization of its N-linked oligosaccharides', *Glycobiology*. Narnia, 10(3), pp. 231–235. doi: 10.1093/glycob/10.3.231.

Medzihradszky, K. F. and Chalkley, R. J. (2015) 'Lessons in de novo peptide sequencing by tandem mass spectrometry.', *Mass spectrometry reviews*. NIH Public Access, 34(1), pp. 43–63.

Meeks, J. P., Arnson, H. A. and Holy, T. E. (2010) 'Representation and transformation of sensory information in the mouse accessory olfactory system', *Nature Neuroscience*. Nature Publishing Group, 13(6), pp. 723–730. doi: 10.1038/nn.2546.

Melrose, D. R., Reed, H. C. B. and Patterson, R. L. S. (1971) 'Androgen Steroids Associated with Boar Odour as an aid to the Detection of Oestrus in Pig Artificial Insemination', *British Veterinary Journal*. W.B. Saunders, 127(10), pp. 497–502. doi: 10.1016/S0007-1935(17)37337-2.

Mills, M. G. L., Gorman, M. L. and Mills, M. E. J. (1980) 'The Scent Marking Behaviour of the Brown Hyaena *Hyaena Brunnea*', *South African Journal of Zoology*. Routledge, 15(4), pp. 240–248. doi: 10.1080/02541858.1980.11447718.

Misawa, K. and Ootsuki, R. (2015) 'PAFFT: A new homology search algorithm for third-generation sequencers', *Genomics*. Academic Press, 106(5), pp. 265–267. doi: 10.1016/J.YGENO.2015.09.005.

Miyazaki, M. *et al.* (2003) 'Molecular cloning and characterization of a novel carboxylesterase-like protein that is physiologically present at high concentrations in the urine of domestic cats (*Felis catus*).', *The Biochemical journal*. Portland Press Limited, 370(Pt 1), pp. 101–10. doi: 10.1042/BJ20021446.

Miyazaki, M. *et al.* (2008) 'The Biological Function of Cauxin, a Major Urinary Protein of the Domestic Cat (*Felis catus*)', in Hurst, J. L. et al. (eds) *Chemical Signals in Vertebrates 11*. New York, NY: Springer, pp. 51–60. doi: 10.1007/978-0-387-73945-8_4.

Miyazaki, M. *et al.* (2018) 'The Chemical Basis of Species, Sex, and Individual Recognition Using Feces in the Domestic Cat', *Journal of Chemical Ecology*. Springer, 44(4), pp. 364–373. doi: 10.1007/s10886-018-0951-3.

Morris, R. S. and Pfeiffer, D. U. (1995) 'Directions and issues in bovine tuberculosis epidemiology and

control in New Zealand', *New Zealand Veterinary Journal*, 43(7), pp. 256–265. doi: 10.1080/00480169./1995.35904.

Morrow, B. A. *et al.* (2000) 'The predator odor, TMT, displays a unique, stress-like pattern of dopaminergic and endocrinological activation in the rat', *Brain Research*. Elsevier, 864(1), pp. 146–151. doi: 10.1016/S0006-8993(00)02174-0.

Mucignat-Caretta, C., Cavaggioni, A. and Caretta, A. (2004) 'Male Urinary Chemosignals Differentially Affect Aggressive Behavior in Male Mice', *Journal of Chemical Ecology*. Kluwer Academic Publishers-Plenum Publishers, 30(4), pp. 777–791. doi: 10.1023/B:JOEC.0000028431.29484.d7.

Mudge, J. M. *et al.* (2008) 'Dynamic instability of the major urinary protein gene family revealed by genomic and phenotypic comparisons between C57 and 129 strain mice.', *Genome biology*. BIOMED CENTRAL LTD, 236 GRAYS INN RD, FLOOR 6, LONDON WC1X 8HL, ENGLAND, 9(5), p. R91. doi: 10.1186/gb-2008-9-5-r91.

Mugford, R. A. and Nowell, N. W. (1971) 'The relationship between endocrine status of female opponents and aggressive behaviour of male mice', *Animal Behaviour*. Academic Press, 19(1), pp. 153–155. doi: 10.1016/S0003-3472(71)80150-1.

Muller-Schwarze, D. *et al.* (1976) 'Response to a mammalian pheromone and its geometric isomer', *J. Chem. Ecol*, 2(3), pp. 389–398.

Munger, S. D., Leinders-Zufall, T. and Zufall, F. (2009) 'Subsystem Organization of the Mammalian Sense of Smell', *Annual Review of Physiology*. Annual Reviews, 71(1), pp. 115–140. doi: 10.1146/annurev.physiol.70.113006.100608.

Nara, K. *et al.* (2011) 'A large-scale analysis of odor coding in the olfactory epithelium.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*. Society for Neuroscience, 31(25), pp. 9179–91. doi: 10.1523/JNEUROSCI.1282-11.2011.

Nazarova, G. G., Proskurniak, L. P. and Yuzhik, E. I. (2016) 'The Presence Of Strange Males' Odor Induces Behavioral Responses And Elevated Levels Of Low Molecular Weight Proteins Excreted In The Urine Of Mature Water Vole Males (Arvicola amphibius L)', *Journal of Chemical Ecology*. Springer US, 42(3), pp. 270–276. doi: 10.1007/s10886-016-0683-1.

Nelson, A. C. *et al.* (2015) 'Protein pheromone expression levels predict and respond to the formation of social dominance networks', *Journal of Evolutionary Biology*, 28(6), pp. 1213–1224. doi: 10.1111/jeb.12643.

Newman, D. J. *et al.* (2000) 'Urinary protein and albumin excretion corrected by creatinine and specific gravity', *Clinica Chimica Acta*, 294(1–2), pp. 139–155. doi: 1016/S0009-8981(00)00181-9.

Nicholas, K. R. *et al.* (1987) *A novel whey protein synthesized only in late lactation by the mammary gland from the tammar (*Macropus eugenii*)*, *Biochem. J*.

Nie, Y. *et al.* (2012) 'Giant panda scent-marking strategies in the wild: role of season, sex and marking surface', *Animal Behaviour*, 84(1), pp. 39–44. doi: 10.1016/j.anbehav.2012.03.026.

Niimura, Y. (2009) 'Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents.', *Human genomics*. BioMed Central, 4(2), pp. 107–18. doi: 10.1186/1479-7364-4-2-107.

Nishimura, K. *et al.* (1989) 'Identification of puberty-accelerating pheromones in male mouse urine', *Journal of Experimental Zoology*. John Wiley & Sons, Ltd, 251(3), pp. 300–305. doi: 10.1002/jez.1402510306.

Nodari, F. *et al.* (2008) 'Sulfated Steroids as Natural Ligands of Mouse Pheromone-Sensing Neurons', *Journal of Neuroscience*. Society for Neuroscience, 28(25), pp. 6407–6418. doi: 10.1523/JNEUROSCI.1425-08.2008.

Norrdahl, K. (1995) 'Population cycles in northern small mammals', *Biological Reviews*. John Wiley & Sons, Ltd (10.1111), 70(4), pp. 621–637. doi: 10.1111/j.1469-185X.1995.tb01654.x.

Novotny, M. *et al.* (1985) 'Synthetic pheromones that promote inter-male aggression in mice.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 82(7), pp. 2059–61. doi: 10.1073/pnas.82.7.2059.

Novotny, M. *et al.* (1986) 'Adrenal-mediated endogenous metabolites inhibit puberty in female mice.', *Science*. American Association for the Advancement of Science, 231(4739), pp. 722–5. doi:

10.1126/SCIENCE.3945805.

Novotny, M. V. *et al.* (1999) 'Positive identification of the puberty-accelerating pheromone of the house mouse: the volatile ligands associating with the major urinary protein', *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1432), pp. 2017–2022. doi: 10.1098/rspb.1999.0880.

Novotny, M. V. *et al.* (2007) 'Chemical Identification of MHC-influenced Volatile Compounds in Mouse Urine. I: Quantitative Proportions of Major Chemosignals', *Journal of Chemical Ecology*. Springer-Verlag, 33(2), pp. 417–434. doi: 10.1007/s10886-006-9230-9.

Novotny, M. V *et al.* (1999) 'A unique urinary constituent, 6-hydroxy-6-methyl-3-heptanone, is a pheromone that accelerates puberty in female mice.', *Chemistry & biology*. CELL PRESS, 600 TECHNOLOGY SQUARE, 5TH FLOOR, CAMBRIDGE, MA 02139 USA, 6(6), pp. 377–83. doi: 10.1016/S1074-5521(99)80049-0.

Novotny, M. V (2003) 'Pheromones, binding proteins and receptor responses in rodents.', *Biochemical Society transactions*. Portland Press Limited, 31(Pt 1), pp. 117–22. doi: 10.1042/.

Nowak, R. *et al.* (2000) 'Role of mother-young interactions in the survival of offspring in domestic mammals.', *Reviews of reproduction*, 5(3), pp. 153–63.

Nyby, J. *et al.* (1977) 'Pheromonal regulation of male mouse ultrasonic courtship (*Mus musculus*)', *Animal Behaviour*. Academic Press, 25, pp. 333–341. doi: 10.1016/0003-3472(77)90009-4.

O'Donnell, C., Weston, K. and Monks, J. (2017) 'Impacts of introduced mammalian predators on New Zealand's alpine fauna', *New Zealand Journal of Ecology*, 41(1), pp. 1–22. doi: 10.20417/nzjecol.41.18.

O'Halloran, K. *et al.* (2005) 'Ecotoxicity of sodium fluoroacetate (compound 1080) to soil organisms', *Environmental Toxicology and Chemistry*. Wiley-Blackwell, 24(5), p. 1211. doi: 10.1897/04-424R.1.

O'Riain, M. . and Jarvis, J. U. . (1997) 'Colony member recognition and xenophobia in the naked mole-rat', *Animal Behaviour*. Academic Press, 53(3), pp. 487–498. doi: 10.1006/ANBE.1996.0299.

Olsen, J. V and Mann, M. (2013) 'Status of large-scale analysis of post-translational modifications by mass spectrometry.', *Molecular & Cellular Proteomics*. American Society for Biochemistry and Molecular Biology, 12(12), pp. 3444–52. doi: 10.1074/mcp.O113.034181.

Ong, S.-E. *et al.* (2002) 'Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics', *Molecular & Cellular Proteomics*. American Society for Biochemistry and Molecular Biology, 1(5), pp. 376–386. doi: 10.1074/MCP.M200025-MCP200.

Paonessa, G. *et al.* (1984) 'Transglutaminase-mediated modifications of the rat sperm surface in vitro.', *Science*. American Association for the Advancement of Science, 226(4676), pp. 852–5. doi: 10.1126/science.6149619.

Parker, M., Needham, M. and White, R. (1982) 'Prostatic steroid binding protein: gene duplication and steroid binding', *Nature*. Nature Publishing Group, 298(5869), pp. 92–94. doi: 10.1038/298092a0.

Penn, D. and Potts, W. (1998) 'MHC-disassortative mating preferences reversed by cross-fostering', *Proceedings of the Royal Society B: Biological Sciences*, 265(1403), pp. 1299–1306. doi: 10.1098/rspb.1998.0433.

Penn, Dustin and Potts, W. K. (1998) 'Untrained mice discriminate MHC-determined odors', *Physiology and Behavior*, 64(3), pp. 235–243. doi: 10.1016/S0031-9384(98)00052-3.

Perkins, D. N. *et al.* (1999) 'Probability-based protein identification by searching sequence databases using mass spectrometry data', *Electrophoresis*. John Wiley & Sons, Ltd, 20(18), pp. 3551–3567. doi: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

Pes, D. *et al.* (1992) 'Isolation of two odorant-binding proteins from mouse nasal tissue', *Biochem. Physiol*, 103(4), pp. 101–1017.

Pes, D. and Pelosi, P. (1995) 'Odorant-binding proteins of the mouse', *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 112(3), pp. 471–479. doi: 10.1016/0305-0491(95)00063-1.

Petersen, T. N. *et al.* (2011) 'SignalP 4.0: discriminating signal peptides from transmembrane regions', *Nature Methods*. Nature Publishing Group, 8(10), pp. 785–786. doi: 10.1038/nmeth.1701.

Petrulis, A. (2013) 'Chemosignals, hormones and mammalian reproduction.', *Hormones and behavior*. ACADEMIC PRESS INC ELSEVIER SCIENCE, 525 B ST, STE 1900, SAN DIEGO, CA 92101-4495 USA, 63(5), pp. 723–41. doi: 10.1016/j.yhbeh.2013.03.011.

Phelan, M. M. *et al.* (2014) 'The Structure, Stability and Pheromone Binding of the Male Mouse Protein Sex Pheromone Darcin', *PLoS ONE*. Edited by P. L. Ho. Public Library of Science, 9(10), p. e108415. doi: 10.1371/journal.pone.0108415.

Piotte, C. P. *et al.* (1998) 'Phylogenetic Analysis of Three Lipocalin-Like Proteins Present in the Milk of Trichosurus vulpecula (Phalangeridae, Marsupialia)', *Journal of Molecular Evolution*. Springer-Verlag, 46(3), pp. 361–369. doi: 10.1007/PL00006313.

Poddar-Sarkar, M. (1996) 'The fixative lipid of tiger pheromone', *Journal of Lipid Mediators and Cell Signalling*, 15, pp. 89–101.

Poran, N. S. (1998) 'Vomeronasal organ and its associated structures in the opossum Monodelphis domestica', *Microscopy Research and Technique*, 43(6), pp. 500–510. doi: 10.1002/(SICI)1097-0029(19981215)43:6<500::AID-JEMT3>3.0.CO;2-H.

Porter, R. H. *et al.* (1991) 'Individual olfactory signatures as major determinants of early maternal discrimination in sheep', *Developmental Psychobiology*, 24(3), pp. 151–158. doi: 10.1002/dev.420240302.

Porter, R. H., Tepper, V. J. and White, D. M. (1981) 'Experiential influences on the development of huddling preferences and "sibling" recognition in spiny mice', *Developmental Psychobiology*. John Wiley & Sons, Ltd, 14(4), pp. 375–382. doi: 10.1002/dev.420140410.

Pracy, L. T. (1962) 'Introduction and Liberation of the Opossum (Trichosurus Vulpecula) Into New Zealand', *Information Series (New Zealand Forest Series)*. 2nd edn. Wellington, N.Z. New Zealand Forest Service, 45, p. 28.

Pratt, J. M. *et al.* (2006) 'Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes', *Nature Protocols*. Nature Publishing Group, 1(2), pp. 1029–1043. doi: 10.1038/nprot.2006.129.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2005) 'NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Research*, 33, pp. D501–D504. doi: 10.1093/nar/gki025.

Del Punta, K. *et al.* (2002) 'Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes', *Nature*. Nature Publishing Group, 419(6902), pp. 70–74. doi: 10.1038/nature00955.

PyMOL (no date) 'The PyMOL Molecular Graphic System'. LLC: Schrödinger.

Rajkumar, R. *et al.* (2010) 'Using mass spectrometry to detect buffalo salivary odorant-binding protein and its post-translational modifications', *Rapid Communications in Mass Spectrometry*, 24(22), pp. 3248–3254. doi: 10.1002/rcm.4766.

Ramm, S. A. *et al.* (2008) 'Comparative Proteomics Reveals Evidence for Evolutionary Diversification of Rodent Seminal Fluid and Its Functional Significance in Sperm Competition', *Molecular Biology and Evolution*. Narnia, 26(1), pp. 189–198. doi: 10.1093/molbev/msn237.

RAMSEY, D. (2005) 'Population dynamics of brushtail possums subject to fertility control', *Journal of Applied Ecology*. Wiley/Blackwell (10.1111), 42(2), pp. 348–360. doi: 10.1111/j.1365-2664.2005.01006.x.

Ranganathan, V. and De, P. K. (1995) 'Androgens and Estrogens Markedly Inhibit Expression of a 20-kDa Major Protein in Hamster Exorbital Lacrimal Gland', *Biochemical and Biophysical Research Communications*. Academic Press, 208(1), pp. 412–417. doi: 10.1006/BBRC.1995.1353.

Ranganathan, V., Jana, N. R. and De, P. K. (1999) 'Hormonal effects on hamster lacrimal gland female-specific major 20 kDa secretory protein and its immunological similarity with submandibular gland major male-specific proteins', *The Journal of Steroid Biochemistry and Molecular Biology*. Pergamon, 70(4–6), pp. 151–158. doi: 10.1016/S0960-0760(99)00103-X.

Rasmussen, L. E. . *et al.* (1996) 'Insect pheromone in elephants', *Nature*, 379(6567), p. 684. doi: 10.1038/379684a0.

Rasmussen, L. E. L. *et al.* (1997) *Purification, Identification, Concentration and Bioactivity of (Z)-7-Dodecen-1-yl Acetate: Sex Pheromone of the Female Asian Elephant, Elephas maxim us*, *Chem.*

Senses.

Rich, T. J. and Hurst, J. L. (1999) *The competing countermarks hypothesis: reliable assessment of competitive ability by potential mates*, *Animal Behaviour*.

Roberts, S. A. *et al.* (2012) 'Pheromonal induction of spatial learning in mice.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 338(6113), pp. 1462–5. doi: 10.1126/science.1225638.

Roberts, S. A. *et al.* (2014) 'Female attraction to male scent and associative learning: the house mouse as a mammalian model', *Animal Behaviour*. Academic Press, 97, pp. 313–321. doi: 10.1016/J.ANBEHAV.2014.08.010.

Roberts, S. A. *et al.* (2018) 'Individual odour signatures that mice learn are shaped by involatile major urinary proteins (MUPs)', *BMC Biology*. BioMed Central, 16(1), p. 48. doi: 10.1186/s12915-018-0512-9.

Roberts, S. C. (2007) 'Scent marking', in Wolff, J. O. and Sherman, P. W. (eds) *Rodent societies: an ecological & evolutionary perspective*. Chicago, Illinois: University of Chicago Press, pp. 255–266.

Roberts, S. C. and Dunbar, R. I. M. (2000) 'Female territoriality and the function of scent-marking in a monogamous antelope (Oreotragus oreotragus)', *Behavioral Ecology and Sociobiology*. Springer-Verlag, 47(6), pp. 417–423. doi: 10.1007/s002650050685.

Robertson, D. H., Beynon, R. J. and Evershed, R. P. (1993) 'Extraction, characterization, and binding analysis of two pheromonally active ligands associated with major urinary protein of house mouse (Mus musculus).', *Journal of chemical ecology*. PLENUM PUBL CORP, 233 SPRING ST, NEW YORK, NY 10013, 19(7), pp. 1405–16. doi: 10.1007/BF00984885.

Robertson, D. H. L. *et al.* (1996) 'Molecular heterogeneity in the Major Urinary Proteins of the house mouse *Mus musculus*', *Biochemical Journal*. Portland Press Limited, 316(1), pp. 265–272. doi: 10.1042/bj3160265.

Robertson, D. H. L. *et al.* (1997) 'Molecular Heterogeneity of Urinary Proteins in Wild House Mouse Populations', *Rapid Communications in Mass Spectrometry*. John Wiley & Sons, Ltd, 11(7), pp. 786–790. doi: 10.1002/(SICI)1097-0231(19970422)11:7<786::AID-RCM876>3.0.CO;2-8.

Robertson, D. H. L. *et al.* (2007) 'Characterization and Comparison of Major Urinary Proteins from the House Mouse, *Mus musculus domesticus*, and the Aboriginal Mouse, *Mus macedonicus*', *Journal of Chemical Ecology*. Kluwer Academic Publishers-Plenum Publishers, 33(3), pp. 613–630. doi: 10.1007/s10886-006-9247-0.

Russell, E. M. (1985) 'The metatherians: order Marsupialia', in Brown, R. E. and MacDonald, D. W. (eds) *Social odours in mammals*. Oxford: Clarendon Press, pp. 45–104.

Sakurai, T. *et al.* (1984) 'PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data', *Biological Mass Spectrometry*. John Wiley & Sons, Ltd, 11(8), pp. 396–399. doi: 10.1002/bms.1200110806.

Sales, G., Czuchnowski, R. and Kapusta, J. (2007) 'Aggression and vocalization behaviour of three sympatric vole species during conspecific and heterospecific same-sex encounters', *Behaviour*. BRILL ACADEMIC PUBLISHERS, PLANTIJNSTRAAT 2, P O BOX 9000, 2300 PA LEIDEN, NETHERLANDS, 144(3), pp. 283–305. doi: 10.1163/156853907780425730.

Šali, A. and Blundell, T. L. (1993) 'Comparative Protein Modelling by Satisfaction of Spatial Restraints', *Journal of Molecular Biology*. Academic Press, 234(3), pp. 779–815. doi: 10.1006/JMBI.1993.1626.

Saltzman, W., Digby, L. J. and Abbott, D. H. (2009) 'Reproductive skew in female common marmosets: what can proximate mechanisms tell us about ultimate causes?', *Proceedings of the Royal Society B: Biological Sciences*. The Royal SocietyLondon, 276(1656), pp. 389–399. doi: 10.1098/rspb.2008.1374.

Samy, E. T. *et al.* (2000) 'Sertoli Cell Prostaglandin D2 Synthetase Is a Multifunctional Molecule: Its Expression and Regulation1', *Endocrinology*. Narnia, 141(2), pp. 710–721. doi: 10.1210/endo.141.2.7329.

Satoh, T. *et al.* (2007) 'Structural basis for recognition of high mannose type glycoproteins by mammalian transport lectin VIP36.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 282(38), pp. 28246–55. doi: 10.1074/jbc.M703064200.

Scaloni, A. *et al.* (2001) 'Purification, cloning and characterisation of odorant- and pheromone-binding proteins from pig nasal epithelium.', *Cellular and molecular life sciences : CMLS*, 58(5–6), pp. 823–34.

Schaal, B. *et al.* (2003) 'Chemical and behavioural characterization of the rabbit mammary pheromone', *Nature*. Nature Publishing Group, 424(6944), pp. 68–72. doi: 10.1038/nature01739.

Schaal, B. *et al.* (2009) 'Mammary olfactory signalisation in females and odor processing in neonates: Ways evolved by rabbits and humans', *Behavioural Brain Research*. Elsevier, 200(2), pp. 346–358. doi: 10.1016/j.bbr.2009.02.008.

Schneider, N. Y. *et al.* (2008) 'The vomeronasal organ of the tammar wallaby', *Journal of Anatomy*, 213(2), pp. 93–105. doi: 10.1111/j.1469-7580.2008.00933.x.

Schwanhäusser, B. *et al.* (2011) 'Global quantification of mammalian gene expression control', *Nature*. Nature Publishing Group, 473(7347), pp. 337–342. doi: 10.1038/nature10098.

Scordato, E. S. and Drea, C. M. (2007) 'Scents and sensibility: information content of olfactory signals in the ringtailed lemur, Lemur catta', *Animal Behaviour*. ACADEMIC PRESS LTD- ELSEVIER SCIENCE LTD, 24-28 OVAL RD, LONDON NW1 7DX, ENGLAND, 73(2), pp. 301–314. doi: 10.1016/j.anbehav.2006.08.006.

Serafini-Cessi, F., Malagolini, N. and Cavallone, D. (2003) 'Tamm-Horsfall glycoprotein: biology and clinical relevance', *American Journal of Kidney Diseases*. W.B. Saunders, 42(4), pp. 658–676. doi: 10.1016/S0272-6386(03)00829-1.

Sheehan, M. J. *et al.* (2016) 'Selection on Coding and Regulatory Variation Maintains Individuality in Major Urinary Protein Scent Marks in Wild Mice', *PLOS Genetics*. Edited by G. S. Barsh, 12(3), p. e1005891. doi: 10.1371/journal.pgen.1005891.

Shen, M.-Y. and Sali, A. (2006) 'Statistical potential for assessment and prediction of protein structures.', *Protein science : a publication of the Protein Society*. Wiley-Blackwell, 15(11), pp. 2507–24. doi: 10.1110/ps.062416606.

Sherborne, A. L. *et al.* (2007) 'The genetic basis of inbreeding avoidance in house mice.', *Current Biology*. CELL PRESS, 600 TECHNOLOGY SQUARE, 5TH FLOOR, CAMBRIDGE, MA 02139 USA, 17(23), pp. 2061–6. doi: 10.1016/j.cub.2007.10.041.

Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.', *Molecular systems biology*. EMBO Press, 7(1), p. 539. doi: 10.1038/msb.2011.75.

Singer, A. G. *et al.* (1986) 'Purification and analysis of a proteinaceous aphrodisiac pheromone from hamster vaginal discharge', *Journal of Biological Chemistry*. AMER SOC BIOCHEMISTRY MOLECULAR BIOLOGY INC, 9650 ROCKVILLE PIKE, BETHESDA, MD 20814, 261(28), pp. 3323–3326.

Singer, A. G. and Macrides, F. (1990) 'Aphrodisin: pheromone or transducer?', *Chemical Senses*. Narnia, 15(2), pp. 199–203. doi: 10.1093/chemse/15.2.199.

Singh, P. B., Brown, R. E. and Roser, B. (1987) 'MHC antigens in urine as olfactory recognition cues', *Nature*. Nature Publishing Group, 327(6118), pp. 161–164. doi: 10.1038/327161a0.

Sippl, M. J. (1993) 'Recognition of errors in three-dimensional structures of proteins', *Proteins: Structure, Function, and Genetics*. John Wiley & Sons, Ltd, 17(4), pp. 355–362. doi: 10.1002/prot.340170404.

Smith, J. L. D., Mcdougal, C. and Miquellet, D. (1989) 'Scent marking in free-ranging tigers, Panthera tigris', *Animal Behaviour*, 37, pp. 1–10.

Söding, J., Biegert, A. and Lupas, A. N. (2005) 'The HHpred interactive server for protein homology detection and structure prediction.', *Nucleic acids research*. Oxford University Press, 33(Web Server issue), pp. W244-8. doi: 10.1093/nar/gki408.

Soini, H. A. *et al.* (2005) 'Stir Bar Sorptive Extraction: A New Quantitative and Comprehensive Sampling Technique for Determination of Chemical Signal Profiles from Biological Media', *Journal of Chemical Ecology*. Kluwer Academic Publishers-Plenum Publishers, 31(2), pp. 377–392. doi: 10.1007/s10886-005-1347-8.

Song, Y. *et al.* (2013) 'High-resolution comparative modeling with RosettaCM.', *Structure*. NIH Public Access, 21(10), pp. 1735–42. doi: 10.1016/j.str.2013.08.005.

Spehr, M. *et al.* (2006) 'Essential Role of the Main Olfactory System in Social Recognition of Major Histocompatibility Complex Peptide Ligands', *Journal of Neuroscience*. Society for Neuroscience, 26(7), pp. 1961–1970. doi: 10.1523/JNEUROSCI.4939-05.2006.

Spinelli, S. *et al.* (2002) 'Boar salivary lipocalin', *European Journal of Biochemistry*. John Wiley & Sons, Ltd (10.1111), 269(10), pp. 2449–2456. doi: 10.1046/j.1432-1033.2002.02901.x.

Spurr, E. B. and Jolly, S. E. (1999) 'Dominant and subordinate behaviour of captive brushtail possums (*Trichosurus vulpecula*)', *New Zealand Journal of Zoology*. Taylor & Francis Group, 26(4), pp. 263–270. doi: 10.1080/03014223.1999.9518195.

Srinivasan, M. S. and Suren, A. (2018) 'Tracking 1080 (sodium fluoroacetate) in surface and subsurface flows during a rainfall event: a hillslope-scale field study', *Australasian Journal of Water Resources*, 22(1), pp. 71–77. doi: 10.1080/13241583.2018.1452329.

Stockley, P., Bottell, L. and Hurst, J. L. (2013) 'Wake up and smell the conflict: odour signals in female competition', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1631). doi: 10.1098/rstb.2013.0082.

Stoddart, D. M. (1982) 'Demonstration of olfactory discrimination by the short-tailed vole, Microtus agrestis L.', *Animal Behaviour*. Academic Press, 30(1), pp. 293–294. doi: 10.1016/S0003-3472(82)80266-2.

Stopkova, R. *et al.* (2009) 'Multiple roles of secretory lipocalins (Mup, Obp) in mice', *Folia Zoologica*, 58(1), pp. 29–40.

Stopková, R. *et al.* (2010a) 'Novel OBP genes similar to hamster Aphrodisin in the bank vole, Myodes glareolus.', *BMC genomics*. BIOMED CENTRAL LTD, 236 GRAYS INN RD, FLOOR 6, LONDON WC1X 8HL, ENGLAND, 11, p. 45. doi: 10.1186/1471-2164-11-45.

Stopková, R. *et al.* (2010b) 'Novel OBP genes similar to hamster Aphrodisin in the bank vole, Myodes glareolus.', *BMC genomics*. BIOMED CENTRAL LTD, 236 GRAYS INN RD, FLOOR 6, LONDON WC1X 8HL, ENGLAND, 11, p. 45. doi: 10.1186/1471-2164-11-45.

Stopková, R. *et al.* (2014) 'Complementary roles of mouse lipocalins in chemical communication and immunity', *Biochemical Society Transactions*, 42, pp. 893–898. doi: 10.1042/BST20140053.

Sun, L. and Müller-Schwarze, D. (1997) 'Sibling recognition in the beaver: a field test for phenotype matching', *Animal Behaviour*. Academic Press, 54(3), pp. 493–502. doi: 10.1006/ANBE.1996.0440.

Sun, S. *et al.* (2014) 'Inhibition of protein carbamylation in urea solution using ammonium-containing buffers', *Analytical Biochemistry*, 446, pp. 76–81. doi: https://doi.org/10.1016/j.ab.2013.10.024.

Tamura, K. *et al.* (2013) 'MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.', *Molecular biology and evolution*. Oxford University Press, 30(12), pp. 2725–9. doi: 10.1093/molbev/mst197.

Taniike, M. *et al.* (2002) 'Perineuronal oligodendrocytes protect against neuronal apoptosis through the production of lipocalin-type prostaglandin D synthase in a genetic demyelinating model.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*. Society for Neuroscience, 22(12), pp. 4885–96. doi: 10.1523/JNEUROSCI.22-12-04885.2002.

Team, R. C. (2017) 'R: A language and environment for statistical computing.' Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Tegoni, M. *et al.* (2000) 'Mammalian odorant binding proteins', *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*. Elsevier, 1482(1–2), pp. 229–240. doi: 10.1016/S0167-4838(00)00167-9.

Thom, M. D. *et al.* (2008) 'Report The Direct Assessment of Genetic Heterozygosity through Scent in the Mouse', *Current Biology*, 18(8), pp. 619–623. doi: 10.1016/j.cub.2008.03.056.

Thompson, A. *et al.* (2003) 'Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS', *Analytical Chemistry*. American Chemical Society, 75(8), pp. 1895–1904. doi: 10.1021/AC0262560.

Thomson, G. M. (1921) *Wild Life in New Zealand: Part 1 - Mammalia*, *New Zealand Board of Science and Art*. Wellington: Marcus F. Marks.

Timm, D. E. *et al.* (2001) 'Structural basis of pheromone binding to mouse major urinary protein (MUP-I)', *Protein Science*, 10, pp. 997–1004. doi: 10.1110/ps.52201.

Tretter, V., Altmann, F. and März, L. (1991) 'Peptide-N4-(N-acetyl-β-glucosaminyl)asparagine

amidase F cannot release glycans with fucose attached α1 → 3 to the asparagine-linked N-acetylglucosamine residue', *European Journal of Biochemistry*, 199(3), pp. 647–652. doi: 10.1111/j.1432-1033.1991.tb16166.x.

Trott, J. F. *et al.* (2002) 'Expression of novel lipocalin-like milk protein gene is developmentally-regulated during lactation in the tammar wallaby, Macropuseugenii', *Gene*. Elsevier, 283(1–2), pp. 287–297. doi: 10.1016/S0378-1119(01)00883-6.

Turton, M. (2007) *Mass spectrometric characterisation of rodent urinary lipocalins*. University of Liverpool. doi: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.634456.

Turton, M. J. *et al.* (2010) 'Roborovskin, a lipocalin in the urine of the Roborovski hamster, *Phodopus roborovskii*.', *Chemical senses*. OXFORD UNIV PRESS, GREAT CLARENDON ST, OXFORD OX2 6DP, ENGLAND, 35(8), pp. 675–84. doi: 10.1093/chemse/bjq060.

Unsworth, J. (2014) *Identification, characterisation and quantification of proteins used in chemical communication*. University of Liverpool. doi: edsble.634456.

Unsworth, J. *et al.* (2017) 'Characterisation of urinary WFDC12 in small nocturnal basal primates, mouse lemurs (Microcebus spp.)', *Scientific Reports*, 7(1), p. 42940. doi: 10.1038/srep42940.

Vandenbergh, J. G. (1969) 'Male Odor Accelerates Female Sexual Maturation in Mice', *Endocrinology*. Narnia, 84(3), pp. 658–660. doi: 10.1210/endo-84-3-658.

Vandoren, G. *et al.* (1983) 'Different Forms of alpha2u-Globulin in Male and Female Rat Urine', *European Journal of Biochemistry*, 134(1), pp. 175–181. doi: 10.1111/j.1432-1033.1983.tb07548.x.

Verplancke, G., Le Boulengé, E. and Diederich, C. (2011) 'Differential marking, investigation and motor activity in presence of conspecific odours differing on their population of origin in bank voles.', *Proceedings of the International Academy of Ecology and Environmental Sciences*, 1(1), pp. 57–69.

Van Vianen, J. *et al.* (2018) 'The effects of single aerial 1080 possum-control operations on common forest birds in the South Island, New Zealand', *New Zealand Journal of Ecology*, 42(2), pp. 169–178. doi: 10.20417/nzjecol.42.17.

Viitala, J. (1977) 'Social organization in cyclic subarctic populations of the voles *Clethrionomys rufocanus* (Sund.) and *Microtus agrestis* (L.)', *Annales Zoologici Fennici*. Finnish Zoological and Botanical Publishing Board, pp. 53–93. doi: 10.2307/23733674.

Vincent, F. *et al.* (2001) 'Crystal structure of aphrodisin, a sex pheromone from female hamster', *Journal of Molecular Biology*, 305(3), pp. 459–469. doi: 10.1006/jmbi.2000.4241.

Wait, R. *et al.* (2001) 'Proteins of rat serum, urine, and cerebrospinal fluid: VI. Further protein identifications and interstrain comparison', *ELECTROPHORESIS*. John Wiley & Sons, Ltd, 22(14), pp. 3043–3052. doi: 10.1002/1522-2683(200108)22:14<3043::AID-ELPS3043>3.0.CO;2-M.

Wang, Y. *et al.* (2018) 'Superoxide dismutases: Dual roles in controlling ROS damage and regulating ROS signaling.', *The Journal of cell biology*. Rockefeller University Press, 217(6), pp. 1915–1928. doi: 10.1083/jcb.201708007.

Wang, Z. *et al.* (2006) 'Pheromone Detection in Male Mice Depends on Signaling through the Type 3 Adenylyl Cyclase in the Main Olfactory Epithelium', *Journal of Neuroscience*. Society for Neuroscience, 16(3), pp. 909–918. doi: 10.1523/jneurosci.1967-06.2006.

Wang, Z. and Storm, D. R. (2011) 'Maternal behavior is impaired in female mice lacking type 3 adenylyl cyclase.', *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*. Nature Publishing Group, 36(4), pp. 772–81. doi: 10.1038/npp.2010.211.

Ward, G. D. (1984) 'Comparison of trap-and radio-revealed home ranges of the brush-tailed possum (Trichosurus vulpecula Kerr) in New Zealand lowland forest', *New Zealand Journal 0/ Zoology*, 11, pp. 85–92. doi: 10.1080/03014223.1984.10428231.

Waridel, P. *et al.* (2007) 'Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automatedde novo sequencing', *Proteomics*. John Wiley & Sons, Ltd, 7(14), pp. 2318–2329. doi: 10.1002/pmic.200700003.

Waterhouse, A. M. *et al.* (2009) 'Jalview Version 2- a multiple sequence alignment editor and analysis workbench', *Bioinformatics*. Oxford University Press, 25(9), pp. 1189–1191. doi: 10.1093/bioinformatics/btp033.

Watson, R. P. *et al.* (2007) 'Three-dimensional structure and ligand binding properties of trichosurin, a metatherian lipocalin from the milk whey of the common brushtail possum *Trichosurus vulpecula*', *Biochemical Journal*, 408(1), pp. 29–38. doi: 10.1042/BJ20070567.

Whelan, S. and Goldman, N. (2001) 'A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.', *Molecular biology and evolution*, 18(5), pp. 691–9.

Whitten, W. K., Bronson, F. H. and Greenstein, J. A. (1968) 'Estrus-Inducing Pheromone of Male Mice: Transport by Movement of Air', *Science*. American Association for the Advancement of Science, 161(3841), pp. 584–585. doi: 10.1126/science.161.3841.584.

Wildlife Act 1975 (1975). New Zealand. Available at: http://www.legislation.vic.gov.au/ (Accessed: 27 November 2018).

Wildlife Regulations (1999). Tasmania.

Wilkinson, T. S. *et al.* (2011) 'WAP domain proteins as modulators of mucosal immunity.', *Biochemical Society transactions*. PORTLAND PRESS LTD, THIRD FLOOR, EAGLE HOUSE, 16 PROCTER STREET, LONDON WC1V 6 NX, ENGLAND, 39(5), pp. 1409–15. doi: 10.1042/BST0391409.

Willis, C. and Poulin, R. (2000) 'Preference of female rats for the odours of non-parasitised males: the smell of good genes?', *Folia Parasitologica*, 47, pp. 6–10.

Wilm, M. (2011) 'Principles of electrospray ionization.', *Molecular & Cellular Proteomics*. American Society for Biochemistry and Molecular Biology, 10(7), p. M111.009407. doi: 10.1074/mcp.M111.009407.

Wong, J. W. H., Cagney, G. and Cartwright, H. M. (2005) 'SpecAlign--processing and alignment of mass spectra datasets', *Bioinformatics*. Oxford University Press, 21(9), pp. 2088–2090. doi: 10.1093/bioinformatics/bti300.

Wong, J. W. H., Durante, C. and Cartwright, H. M. (2005) 'Application of Fast Fourier Transform Cross-Correlation for the Alignment of Large Chromatographic and Spectral Datasets', *Analytical Chemistry*. American Chemical Society, 77(17), pp. 5655–61. doi: 10.1021/AC050619P.

Woolhouse, A. D., Weston, R. J. and Hamilton, B. H. (1994) 'Analysis of secretions from scent-producing glands of brushtail possum (*Trichosorus vulpecula*, Kerr)', *Journal of Chemical Ecology*, 20(2), pp. 239–253. doi: 10.1007/BF02064434.

Wu, A. H. B. (2002) 'Diagnostic enzymology and other biochemical markers of organ damage', in McClatchey, K. D. (ed.) *Clinical Laboratory Medicine*. 2nd edn. Philadelphia: Lippincott, Williams & Wilkins, pp. 281–321.

Wyatt, T. D. (2014) *Pheromones and Animal Behavior: Chemical Signals and Signatures*. 2nd edn. Cambridge: Cambridge University Press. doi: DOI: 10.1017/CBO9781139030748.

Yamazaki, K. *et al.* (1976) 'Control of mating preferences in mice by genes in the major histocompatibility complex*', *Journal of Experimental Medicine*, 144(5), pp. 1324–1335. doi: 10.1084/jem.144.5.1324.

Yamazaki, K. *et al.* (1983) 'Recognition of H-2 types in relation to the blocking of pregnancy in mice.', *Science*. American Association for the Advancement of Science, 221(4606), pp. 186–8. doi: 10.1126/SCIENCE.6857281.

Yamazaki, K. *et al.* (2000) 'Parent-progeny recognition as a function of MHC odortype identity', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 97(19), pp. 10500–10502. doi: 10.1073/pnas.180320997.

Yang, H. *et al.* (2005) 'Composition and evolution of the V2r vomeronasal receptor gene repertoire in mice and rats', *Genomics*. Academic Press, 86(3), pp. 306–315. doi: 10.1016/J.YGENO.2005.05.012.

Yenugu, S. *et al.* (2004) 'Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif.', *Biology of reproduction*. SOC STUDY REPRODUCTION, 1603 MONROE ST, MADISON, WI 53711-2021 USA, 71(5), pp. 1484–90. doi: 10.1095/biolreprod.104.031567.

Yoshikawa, K. *et al.* (2013) 'An unsaturated aliphatic alcohol as a natural ligand for a mouse odorant receptor', *Nature Chemical Biology*. Nature Publishing Group, 9(3), pp. 160–162. doi: 10.1038/nchembio.1164.

Zabaras, R., Richardson, B. J. and Wyllie, S. G. (2005) 'Evolution in the suite of semiochemicals secreted by the sternal gland of Australian marsupials', *Australian Journal of Zoology*, 53(4), pp. 257–263. doi: 10.1071/ZO04070.

Zhang, J.-X., Zhang, Z.-B. and Wang, Z.-W. (2001) 'Scent, social status, and reproductive condition in rat-like hamsters (Cricetulus triton)', *Physiology & Behaviour*, 74, pp. 415–420.

Zhang, J. and Webb, D. M. (2003) 'Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(14), pp. 8337–41. doi: 10.1073/pnas.1331721100.

Zhang, Xiaohong, Zhang, Xinmin and Firestein, S. (2007) 'Comparative genomics of odorant and pheromone receptor genes in rodents', *Genomics*. Academic Press, 89(4), pp. 441–450. doi: 10.1016/J.YGENO.2007.01.002.

Zhang, Z., Yang, M. J. and Pawliszyn, J. (1994) 'Solid-Phase Microextraction. A Solvent-Free Alternative for Sample Preparation', *Analytical Chemistry*. American Chemical Society, 66(17), pp. 844A-853A. doi: 10.1021/ac00089a001.

Zhokhov, S. S. *et al.* (2017) 'An EThcD-Based Method for Discrimination of Leucine and Isoleucine Residues in Tryptic Peptides', *Journal of The American Society for Mass Spectrometry*. Springer US, 28(8), pp. 1600–1611. doi: 10.1007/s13361-017-1674-3.

Zhu, J. *et al.* (2017) 'Reverse chemical ecology: Olfactory proteins from the giant panda and their interactions with putative pheromones and bamboo volatiles.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(46), pp. E9802–E9810. doi: 10.1073/pnas.1711437114.

An exploration of the protein component of scent marks from a range of mammalian species using a mass spectrometry-based approach

Supplementary Material

By Grace May Loxley

September 2019

Contents

3 Characterisation of the protein content of scent secretions in the bank vole, *Myodes glareolus*. 2
3.6 Supplementary Material 2
3.6.1 Published Paper 2
3.6.2 Individual Intact Mass Spectra 14
3.6.3 Glareosin sequencing fragment ion data (from paper supplementary) 26
3.6.4 Glareosin heavy leucine data 38
3.6.5 Multiple sequence alignment for phylogenetic analysis (from paper supplementary) 40
3.6.6 Bank Vole Bladder Urine Intact Protein Mass Spectra 41
3.6.7 Female scent mark intact protein mass spectra 43
4 Characterisation of the urinary protein content in the field vole, *Microtus agrestis* 44
4.6 Supplementary Material 44
4.6.1 Intact mass profiles of urine from mature field voles 44
4.6.2 Fragment ion spectra of peptides used to sequence field vole glareosin. 52
4.6.3 Sequence coverage of odorant binding proteins from analysis of pooled female field vole urine. 91
4.6.4 Alignment of major urinary proteins to support preliminary sequencing of a field vole major urinary protein 92
4.6.5 Intact mass profiles of juvenile field vole urine. 93
5 Characterisation of the urinary protein content in the New Zealand brushtail possum, *Trichosurus vulpecula* 94
5.6 Supplementary Information 94
5.6.1 Intact mass profiles of individual samples 94
5.6.2 MALDI-ToF peptide mass fingerprints from tryptic in-gel digestion 103
5.6.3 Homologous sequences to the trichosurin-like protein from M. domestica 105
5.6.4 Fragment ion spectra for sequencing 106
5.6.5 Fragment ion spectra for determination of leucine and isoleucine residues in peptides with two ambiguous leucine or isoleucine residues. 132
5.6.6 Table of accession numbers for multiple sequence alignment 133
5.6.7 Multiple sequence alignment of lipocalins 137
6 Characterisation of urinary WFDC12 in nocturnal basal primates, mouse lemurs (*Microcebus spp.*) 138

3 Characterisation of the protein content of scent secretions in the bank vole, *Myodes glareolus*

3.6 Supplementary Material

3.6.1 Published Paper

OPEN BIOLOGY

rsob.royalsocietypublishing.org

Research

Cite this article: Loxley GM, Unsworth J, Turton MJ, Jebb A, Lilley KS, Simpson DM, Rigden DJ, Hurst JL, Beynon RJ. 2017 Glareosin: a novel sexually dimorphic urinary lipocalin in the bank vole, *Myodes glareolus*. *Open Biol.* **7**: 170135.
http://dx.doi.org/10.1098/rsob.170135

Received: 2 June 2017
Accepted: 27 July 2017

Subject Area:
biochemistry/cognition/developmental biology/structural biology

Keywords:
bank vole, *Myodes glareolus*, odorant-binding protein, metabolic labelling, mass spectrometry, glareosin

Author for correspondence:
Robert J. Beynon
e-mail: r.beynon@liverpool.ac.uk

†Present address: Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK.

This paper is dedicated to the memory of Ellen Higginbottom, a lovely young scientist who gained her first experience of mass spectrometry in the Centre for Proteome Research.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3859369.

THE ROYAL SOCIETY PUBLISHING

Glareosin: a novel sexually dimorphic urinary lipocalin in the bank vole, *Myodes glareolus*

Grace M. Loxley[1], Jennifer Unsworth[1], Michael J. Turton[1], Alexandra Jebb[3], Kathryn S. Lilley[2,4,†], Deborah M. Simpson[1], Daniel J. Rigden[1], Jane L. Hurst[3] and Robert J. Beynon[1]

[1]Centre for Proteome Research, Institute of Integrative Biology, and [2]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
[3]Mammalian Behaviour and Evolution Group, Institute of Integrative Biology, University of Liverpool, Leahurst Campus, Neston CH64 7TE, UK
[4]Department of Biochemistry, University of Leicester, Leicester LE1 7RH, UK

GML, 0000-0002-1884-5461; KSL, 0000-0003-0594-6543; DMS, 0000-0002-3962-4895; DJR, 0000-0002-7565-8937; JLH, 0000-0002-3728-9624; RJB, 0000-0003-0857-495X

The urine of bank voles (*Myodes glareolus*) contains substantial quantities of a small protein that is expressed at much higher levels in males than females, and at higher levels in males in the breeding season. This protein was purified and completely sequenced at the protein level by mass spectrometry. Leucine/isoleucine ambiguity was completely resolved by metabolic labelling, monitoring the incorporation of dietary deuterated leucine into specific sites in the protein. The predicted mass of the sequenced protein was exactly consonant with the mass of the protein measured in bank vole urine samples, correcting for the formation of two disulfide bonds. The sequence of the protein revealed that it was a lipocalin related to aphrodisin and other odorant-binding proteins (OBPs), but differed from all OBPs previously described. The pattern of secretion in urine used for scent marking by male bank voles, and the similarity to other lipocalins used as chemical signals in rodents, suggest that this protein plays a role in male sexual and/or competitive communication. We propose the name glareosin for this novel protein to reflect the origin of the protein and to emphasize the distinction from known OBPs.

1. Background

Olfactory communication is prevalent in rodents, where semiochemicals are capable of transmitting information regarding identity, relatedness, territory, health status and mating availability [1–5]. Chemosignalling is highly conserved, with many species displaying scent marking behaviours that make use of urine, faeces and glandular secretions to convey information. Members of the lipocalin protein family are often involved in chemosignalling, and are found in several rodent secretions and tissues where they serve this role, including nasal tissue, saliva, urine, tears and vaginal discharge [6–9]. Murine rodents (Old World rats and mice, sub-family Murinae) express a set of proteins known as major urinary proteins (MUPs), which can be highly polymorphic, whereas hamsters and voles (sub-families Cricetinae and Arvicolinae) seem to express chemosignalling lipocalins more typical of the odorant-binding protein (OBP) family.

Urinary protein expression has been well characterized in the house mouse (*Mus musculus domesticus*). The highly polymorphic MUPs, expressed by both males and females, can communicate individual identity, kinship, dominance, and potentially oestrus and health status [10–19]. Sexual dimorphism is pronounced, with males typically expressing three- to fourfold more MUPs overall than females, while some MUPs are expressed almost exclusively by males [6].

MUPs mediate chemosignalling, either by direct detection through vomeronasal 2 (V2R) receptors in the vomeronasal organ [11] or by binding volatile components, promoting their slow release over a prolonged period, and extend the lifespan of the scent mark [13,20,21]. The protein complement of rat urine also has a similarly polymorphic expression of homologous MUPs, but with much stronger sexual dimorphism [22,23].

Relatively little is yet known about the expression of chemosignalling proteins in the vole family, but sexual dimorphism in urinary protein expression has been observed in the bank vole, *Myodes glareolus*, where protein levels are much higher in males [24]. Bank voles live in small, mixed-sex groups during the winter that break up in the breeding season. While breeding females inhabit non-overlapping home ranges close to the over-wintering site, males have larger overlapping home ranges within hierarchical groups that overlap several females [25,26]. Males deposit urine around their territories in numerous small scent marks, using long brush-like hairs on the prepuce tip to streak out their scent [27,28], contrasting with the excretion of urine in pools by females [28,29]. Scent marking rates are particularly high in new environments, while dominant males also mark subordinate male burrow and nest areas continually. Females prefer males that scent mark more frequently [30]. Three male-specific OBPs have been identified in male bank vole urine that might play a role in chemical signalling [31]. To understand the expression and potential role of urinary proteins in bank vole communication further, we examined the expression of urinary proteins in wild-caught and captive-bred voles in the breeding and non-breeding season. Here, we characterize a new urinary protein in *M. glareolus*, distinct from those previously identified, that is expressed at high level only by males and only in the breeding season. The complete protein sequence was obtained primarily using in-solution protease digestion followed by tandem mass spectrometry, distinguishing between the otherwise isobaric amino acids leucine and isoleucine using metabolic labelling. Homology modelling and structural analysis reveal strong similarity to known OBPs, but this protein is distinct from those previously described in bank voles or in other species and is the most abundant urinary protein expressed by male bank voles. Given the potentially important investment by male bank voles in this particular urinary protein during the breeding season, we propose the name glareosin to distinguish this from other OBPs.

2. Material and methods

2.1. Sampling

Urine samples were collected from both wild-caught and captive-bred *M. glareolus* voles derived from two different geographical areas of the UK (Wirral Peninsula in Merseyside, approx. 53.288° N, −3.028° E, and Kielder Forest in Northumberland, approx. 55.208° N, −2.528° E). Urine was freely expressed and collected from clean plastic cages prior to measurement of protein and creatinine concentration.

2.2. In-gel proteolysis

Protein bands from SDS–PAGE were digested with trypsin to generate peptides suitable for further analysis by MALDI-ToF mass spectrometry. Excised gel plugs (approx. 1 mm^3) were destained, then reduced and carbamidomethylated. Peptides were recovered for mass spectrometric analysis.

2.3. Edman degradation

SDS–PAGE gels before staining were electroblotted to polyvinylidene difluoride (PVDF) membranes for N-terminal sequencing using an Applied Biosystems 476A gas-phase sequencer (Applied Biosystems). After electroblotting, the PVDF was stained with Coomassie blue to visualize protein bands prior to excision and Edman degradation.

2.4. MALDI-ToF mass spectrometry

Analysis of peptides from in-gel digests was undertaken using a MALDI-ToF reflectron mass spectrometer (Waters, Manchester, UK) in positive ion mode. All aspects of data acquisition, processing and machine management were controlled through the MassLynx software suite (v. 4.0).

2.5. In-solution proteolysis

Aliquots of protein (10 µg) purified from cage deposits by anion exchange chromatography were reduced, alkylated and digested with trypsin, endopeptidase GluC or endopeptidase LysC.

2.6. Electrospray ionization mass spectrometry

Electrospray ionization mass spectrometry (ESI-MS) was used in two modes: liquid chromatography–mass spectrometry (LC–MS) was used for intact mass analysis while tandem mass spectrometry (MS/MS) was used for peptide sequence analysis. All ESI-MS was undertaken on a Q-ToF Micro mass spectrometer (Waters, Manchester, UK) in positive ion mode. As an additional aid in the interpretation of tandem mass spectra, peptides were isotopically labelled with ^{18}O by performing proteolytic digestion in a 1 : 1 mixture of light ($H_2[^{16}O]$) and heavy ($H_2[^{18}O]$) water. Incorporation of a 1 : 1 mixture of $[^{16}O]$ and $[^{18}O]$ atoms into the newly formed C-termini of peptides prior to tandem mass spectrometry allowed y-ions to be identified as a sequence of doublets of approximately equal intensity, separated by 2 Da. To confirm and complete the sequence, we repeated the digestions and analysed the samples on a high-resolution instrument with high mass accuracy and resolution for precursor and product ions. For this stage, samples were analysed using a Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific, Hemel Hempstead, UK) coupled with a QExactive mass spectrometer (Thermo Fisher Scientific).

2.7. Use of labelled dietary leucine to discriminate isoleucine from leucine

To discriminate between isobaric leucine and isoleucine residues, we fed bank voles a diet containing stable isotope-labelled leucine. Cage-deposited urine samples were collected from four voles (day 0) before they were transferred to a new cage with the $[^2H_3]$ leucine diet provided *ad libitum*. Urinary proteins were reduced, alkylated and digested with trypsin in solution, followed by LC–MS/MS analysis on the QExactive-HF (Thermo Scientific) as described above. Leucine and

isoleucine residues were then manually assigned from the raw data and confirmed with MASCOT and PEAKS searches under the same search conditions as below with triple labelling with deuterium as an additional variable modification, against the derived sequence of glareosin.

2.8. Protein sequence analysis

The final amino acid sequence was used in a BLAST search [33] using default parameters for protein matches against Rodenta. The 138 matches were reduced and processed as follows. First, incomplete sequences, sequences substantially larger than the core lipocalin size of approx. 160 amino acids or those that only matched across part of the sequence were eliminated. Some sequence entries were exact duplicates and were reduced to single entries. Finally, because we wished to compare the glareosin-secreted protein sequence, signal peptides were removed, either guided by the feature entry in the database entry or through the SignalP 4.1 server [34] (http://www.cbs.dtu.dk/services/SignalP/). The reduced sequence set was aligned with MAFFT using the high accuracy linsi algorithm [35] with JALVIEW [36] used to display and manipulate sequence alignments.

2.9. Phylogenetic analysis

The evolutionary history was inferred by using the maximum-likelihood method based on the JTT matrix-based model [37]. Bootstrapping analysis [38] using 500 replicates was carried out. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. All positions containing gaps and missing data were eliminated leaving a total of 112 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [39].

2.10. Homology modelling

The structure of mature glareosin, without its signal peptide, was modelled using the RosettaCM protocol [40]. Ten models were produced for each combination of templates and alignments. Templates were identified from a non-redundant library of PDB structures using the HHpred server [41], and modelling was done with one, five or 10 templates assessing the results quantitatively with Rosetta's own energy function and with the Prosa II [42], DOPE [43] and QMEAN [44] protein structure quality metrics. Stereochemistry was assessed with PROCHECK [45]. Structures were superimposed using GESAMT [46]. Cavities were detected and measured using the GHECOM [47] and Profunc [48] servers. PyMOL (https://www.pymol.org/) was used to visualize and manipulate structures and to produce structure figures.

3. Results and discussion

To assess seasonal and sex variation in urinary protein output, urine samples were obtained from wild-caught bank voles captured during the breeding and non-breeding season (to correct for differences in urine dilution, protein output was expressed as mg (mg creatinine)$^{-1}$). These analyses confirmed that urinary protein output was substantially higher in males, but only during the breeding season (interaction between season and sex, $F_{1,21} = 5.19$, $p = 0.033$; figure 1*a*). Male average protein output increased over threefold, from 3.5 ± 0.5 mg protein (mg creatinine)$^{-1}$ during the non-breeding season (uncorrected urinary protein concentration 0.36 ± 0.07 mg ml^{-1}) up to 11.2 ± 1.2 mg protein (mg creatinine)$^{-1}$ in the breeding season (uncorrected urinary protein concentration 1.76 ± 0.27 mg ml^{-1}). As urinary creatinine levels were not influenced by season or sex, these differences in urinary protein output were entirely due to differences in the concentration of protein excreted in urine. A preliminary assessment of protein complexity in these samples by one-dimensional (1D) SDS–PAGE revealed an intense band between 14 and 21 kDa that was evident only in male samples and only during the breeding season (figure 1*b*,*c*). We also assessed protein output in urine samples from bank voles bred in captivity and kept under breeding season lighting conditions but without sexual experience. This confirmed a highly significant sex difference in urine protein output ($F_{1,28} = 79.6$, $p < 0.0001$), with levels comparable to those seen in wild-caught voles during the breeding season (figure 1*a*). Thus, elevated protein output in males was not dependent on sexual experience. This elevated protein output was evident in male bladder urine, sampled when older voles were culled (effect of sex, $F_{1,10} = 6.8$, $p = 0.026$). SDS–PAGE confirmed that the same intense band between 14 and 21 kDa was present in male but not in female samples, in both naturally deposited and bladder urine (data not shown).

Intact mass analysis has also been used to assess the heterogeneity of urinary proteins in both captive-bred [49–52] and wild-caught mice [10,17,53,54], identifying small mass changes caused by discrete amino acid substitutions in the protein sequence. The intact mass profile of the *M. glareolus* urinary proteins was analysed by ESI-MS (figure 1*d*). A single predominant intact mass was measured in all samples at $16\,930 \pm 1$ Da, and there was no evidence of inter- or intra-individual heterogeneity in the mass profile. The protein identified at a mass of 16 930 Da in all *M. glareolus* urine samples was purified by anion exchange chromatography. This ion exchange purified protein was recovered and used for primary sequence analysis, as genomic or transcriptomic data were lacking. The measured intact mass differed from the predicted masses of the three urinary OBP proteins reported in bank voles by Stopková *et al*. [31], which, allowing for the formation of two putative disulfide bonds, together with loss of signal peptide predicted by signalP [34], were OBP1 (D3VW62_MYOGA): 16 643 Da, OBP2 (D3VW64_MYOGA): 16 837 Da and OBP3 (D3VW62_MYOGA): 16 749 Da, consistent with this being a novel protein.

After 1D SDS–PAGE and blotting to PVDF membrane, the 16 930 Da protein was partially sequenced by gas-phase Edman degradation. Although less commonly used today, Edman degradation permits precise positioning of the true N-terminal sequence of the protein. The recovered sequence HSEIDEKWVTVAIAADNVNK used in searching (BlastP) [55] with standard search parameters aligned most strongly to the N-terminal sequence of a prairie vole (*Microtus ochrogaster*) aphrodisin-like protein 1 (best match: XP_005372052; 70% identity) and a bank vole (*M. glareolus*) OBP1 (best match, D3VW62_MYOGA; 65% identity) as well as other members of the lipocalin family. This match pointed to the potential role of this urinary protein as a semiochemical lipocalin. Although the N-terminal sequence overlapped with the first structurally conserved GlyXxxTrp region of the lipocalin family (GXW [56]), the highly conserved glycine residue of the motif was absent. However, Glu (E) and Gly (G) elute in

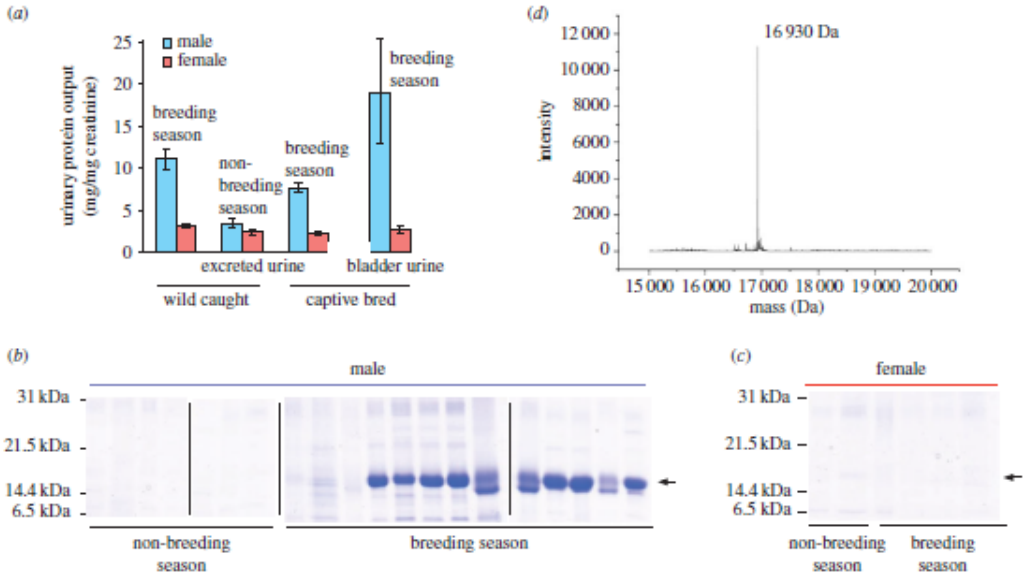


Figure 1. Analysis of bank vole urinary protein output. (*a*) For urine samples from adult male or female bank voles (see text), protein and creatinine concentrations were determined and expressed as mg protein (mg creatinine)$^{-1}$ to correct for urine dilution. The protein complement was analysed by SDS–PAGE for (*b*) male and (*c*) female voles (vertical bars separate discrete gels) and by (*d*) electrospray ionization mass spectrometry (male).

close proximity in Edman degradation, raising the possibility of a mis-call at this position.

To gain further information about the 16 930 Da protein, a peptide sequencing strategy based on mass spectrometry was adopted. Two approaches were taken. First, Q-TOF tandem mass spectrometry of peptides obtained by direct infusion of proteolytic digests of the purified protein and secondly, LC–MS/MS of the peptide mixture on a second instrument that generated product ions at high mass accuracy and resolution. The sequencing strategy was based on digestion with three different endopeptidases (trypsin, endopeptidase LysC and endopeptidase GluC) to generate overlapping peptides that would cover as much of the primary sequence of the mature protein as possible, although unable to discriminate between the isobaric Leu/Ile pair, signified here by the residue 'J'. In some instances, interpretation of the fragment ion mass spectra was assisted by labelling peptides using a 1 : 1 ratio of $H_2^{16}O : H_2^{18}O$ in the digestion reaction. Only the y-series of ions, derived from the C-terminus of each peptide, are isotopically labelled in this reaction, and the doublets thus facilitated discrimination of the b- and y-ion series. Following interpretation of the amino acid sequence from the fragmented peptide, the theoretical m/z value of the $[M+H]^+$ peptide was calculated and reconciled with the ions observed by MALDI-ToF. The complete sequence strategy is presented in figure 2, and the relevant peptide mass spectra are presented in the electronic supplementary material.

Edman degradation predicted an N-terminal tryptic peptide (HSEIDEK) with a theoretical $[M+H]^+$ mass of $m/z = 857.4$. No peptide was detected at $[M+H]^+$ 857.4 Da in either MALDI-ToF MS analysis of trypsin or LysC peptides. However, fragmentation of the tryptic peptide at $[M+2H]^{2+}$ $m/z = 393.21$ yielded the sequence HSEJDGK (figure 2, peptide t1). This sequence included the highly conserved glycine residue of the N-terminal lipocalin motif (GXW), aligned with the ambiguous G/E call from the Edman sequencing confirming a glycine residue at this position. The second tryptic peptide within the Edman sequence was predicted as $[M+2H]^{2+}$, $m/z = 700.9$; the sequence was determined as WVTVAJAADNVNK (t2) from the b- and y-ion series using ^{18}O labelling; this contained the tryptophan residue of the GXW conserved motif. The N-terminal region was extended by tandem MS of a miscleaved peptide $[M+3H]^{3+}$, $m/z = 748.08$ as WVTVAJAADNVNKJEEGGSJR (t3), also present at $[M+H]^+$ $m/z = 2242.13$ in MALDI-ToF MS analysis of tryptic peptides. The sequence of the $[M+H]^+$ 2242.13 tryptic peptide was confirmed by the $[M+2H]^{2+}$ 430.7 tryptic peptide t4 (JEEGGSJR).

Since a feature of OBP-like proteins is the presence of two conserved disulfide bonds, the positions of cysteine residues were identified by carbamidomethylation. MALDI-ToF analysis of tryptic peptides from non-reduced preparations identified two peptides at $[M+H]^+$ $m/z = 1137.51$ and $m/z = 2131.04$ that were shifted upon carbamidomethylation to $[M+H]^+$ $m/z = 1253.56$ and $m/z = 2247.10$, a Δmass of 116 Da. The sequence of the reduced and alkylated peptide $[M+H]^+$ $m/z = 1253.56$, isolated on LC–MS as the $[M+2H]^{2+}$ $m/z = 627.25$ (T5), was EJTC*TEAC*NR, containing two modified cysteine residues. The Δmass of 116 Da following reduction and alkylation could not be explained simply by the carbamidomethylation of the two cysteine residues, which would generate a Δmass of 114.032 Da (2 × 57.016 Da). The additional 2 Da difference is explained by the reduction of a disulfide bond formed between the two cysteine residues. Since the unmodified peptide $[M+H]^+$ $m/z = 1137.51$ is detected in oxidizing conditions, neither cysteine residue could have formed a disulfide bond with a second cysteine residue from a different region of the protein.

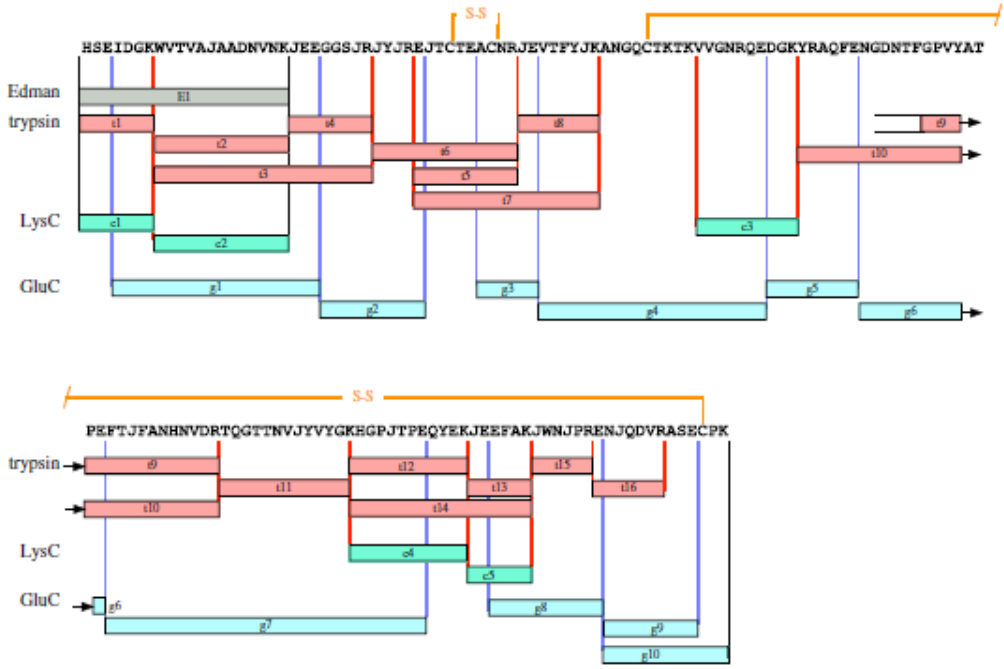


Figure 2. Complete amino sequence of the novel bank vole urinary protein. The bank vole urinary protein was digested with multiple endopeptidases (t, trypsin; c, endopeptidase LysC; g, endopeptidase GluC) and sequenced *de novo* by tandem mass spectrometry. In addition, the Edman degradation data of the intact protein allowed definition of the true N-terminus. The symbol 'J' is used to highlight the ambiguity between leucine and isoleucine in all positions other than the N-terminus, where Edman degradation was unambiguous. The positions of the disulfide bond are inferred by homology with similar proteins.

Furthermore, a high-resolution peptide T6 $[M + 2H]^{2+}$ m/z = 842.92 sequenced as JYJREJTCTEAC*NR. A tight disulfide loop separated by three amino acids is also a feature of other lipocalins and OBPs, including aphrodisin [57]; this provided further presumptive evidence that this protein is an aphrodisin-like lipocalin.

Using similar logic and further tandem MS, the entire sequence of the protein was recovered. All high-resolution peptide tandem mass spectra and sequence calls are provided in the electronic supplementary material. The protein sequence predicted a total length of 149 amino acids. The predicted average mass of the protein was 16 934 Da, which, when adjusted to 16 930 Da to allow for the loss of 4 Da through formation of the disulfide bonds at C_{36}–C_{40} (proved) and C_{55}–C_{147} (surmised, but consistent with homology modelling), correctly predicted the intact mass measured for the urinary protein.

Mass spectrometry-based sequencing *de novo* cannot distinguish between the isobaric amino acids leucine and isoleucine. To discriminate between this isobaric pair, voles were fed a diet partially labelled (relative isotope abundance of approx. 0.5) with [$^{2}H_3$]leucine. Because the protein was secreted in the urine, we surmised that the incorporation of this essential amino acid would result in specific labelling of leucine residues in the protein and in peptides derived therefrom. Both leucine and isoleucine are essential amino acids, and there is no mammalian metabolic pathway whereby the labelling centres in leucine could be transferred to isoleucine. After digestion with trypsin or endopeptidase Glu-C, or a double digest using both endopeptidases, partial labelling meant that each peptide (of monoisotopic mass M) containing a single leucine residue would be accompanied by a second mass, 3 Da heavier, leading to an M, M+3 Da doublet in both precursor and product ion spectra. Peptides containing solely isoleucine residues would not show any labelling doublet. Finally, peptides containing more than one leucine/isoleucine residue would require further analysis to locate the position of the leucine residues. The strategy is illustrated in figure 3, together with labelling profiles for several urinary glareosin peptides.

All leucine/isoleucine ambiguities were evaluated manually and assigned from the raw data (figure 4). Tryptic peptides containing a single ambiguous site (defined as 'J'; HSEIDGK isobaric identity known from Edman degradation, WVTVAJAADNVNK, EJTCTEACNR, TQGTTNVJYVYGK, HGPJTPEQYEK, ENJQDVR, ACNRJLE, VTFYJK, FTJFANH NVDR) were readily resolved from the precursor ion spectra. For peptides that contained more than a single instance of leucine or isoleucine, the strategy was more complicated. Most simply, precursor ion spectra could disambiguate peptides that contained two of the same residues (JWNJIPR), as the mass shift was unambiguous (+6 Da, Leu/Leu; 0 Da, Ile/Ile). For peptides that contained two Leu/Ile residues, only one of which was labelled, the precursor ion mass shift indicated the number but not the position of the leucine and isoleucine residues. Positional resolution was achieved by inspection of fragment ion spectra (electronic supplementary material). Fragment ion spectra were examined for+3 Da increases in the y- and b-ion series at each Leu/Ile

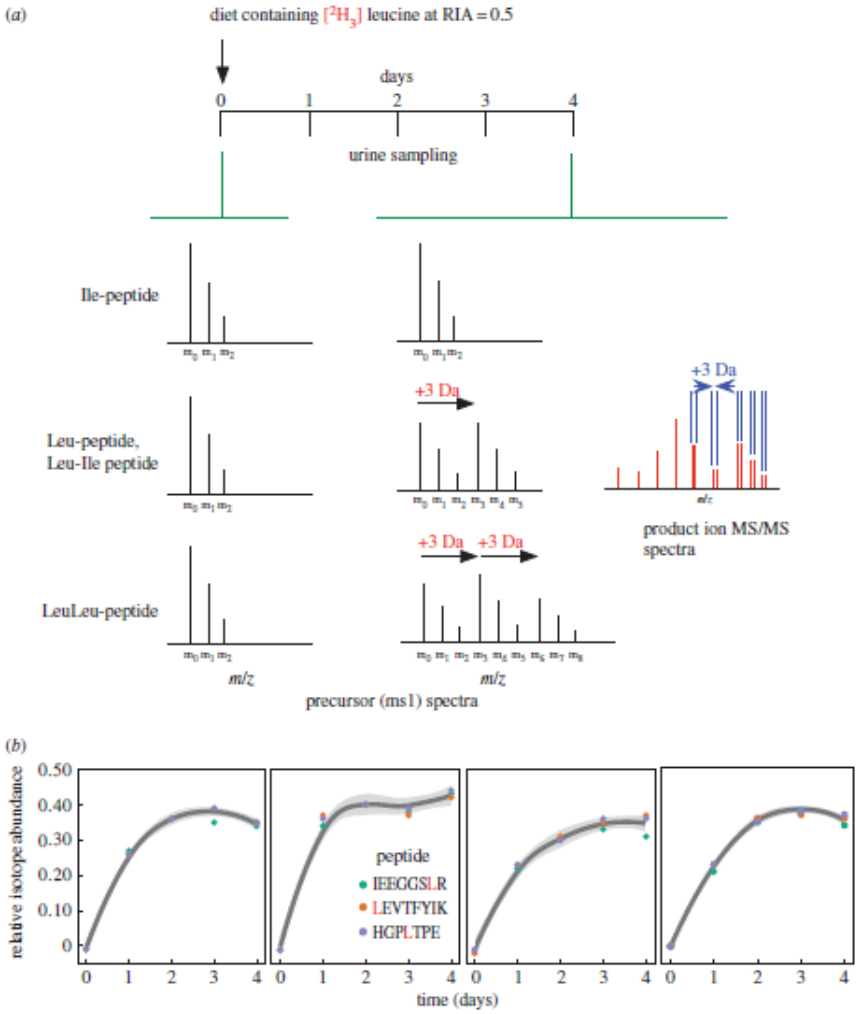


Figure 3. Metabolic labelling strategy to discern leucine and isoleucine residues. (*a*) Bank voles were fed a diet containing [$^2H_3$] leucine at a relative isotope distribution of 0.5. Incorporation of stable isotope-labelled leucine into peptides permits discrimination between leu and ile residues, either from precursor or product ion mass spectra. (*b*) After 4 days of labelling, urinary proteins (three representative peptides are show here for four voles) were labelled to the same extent as the dietary precursor and used for sequencing *de novo*.

product ion, clearly flagged as a doublet. This defined the position of Leu and Ile for peptides JEEGGSJR, JYJRE and JVTFYJK (electronic supplementary material). The remaining unassigned Leu/Ile site was in the small tryptic peptide JEEFAK (t13; $[M+H]^+ = 736.3875$ *m/z*), which is identical to an equivalent tryptic peptide derived from OBP2 and OBP3 [31]. To resolve this issue, we assigned the residue identity using tryptic missed cleavage peptides (this work: HGPLTPEQYEKJEEFAK compared to the peptide GQPLTPEQYEKLEEFAK from OBP2 and OBP3 (Uniprot D3VW63_MYOGA and D3VW64_MYOGA), respectively). The first leucine residue for the protein described here had already been confirmed (see previously) and the precursor mass spectrum of the missed cleavage peptide had a fragment ion distribution consonant with one leucine residue and one isoleucine residue, whereas the OBP peptide also present in LC–MS/MS analysis displayed a fragment isotopic distribution consistent with the presence of two heavy leucine residues (data not shown).

We were thus able to derive the complete, unambiguous sequence of the bank vole urinary protein, including the identification of all leucine and isoleucine residues. The entire sequence was used in a BLAST search against all rodent sequences. The first major conclusion is that this abundant protein in bank vole urine is novel, and has not been reported previously. To distinguish this protein from other bank vole urinary proteins [31], we therefore propose the name 'glareosin' (derived from the species *M. glareolus*). The glareosin sequence matched to several lipocalins, most strongly to aphrodisins and OBPs, with weaker matches to

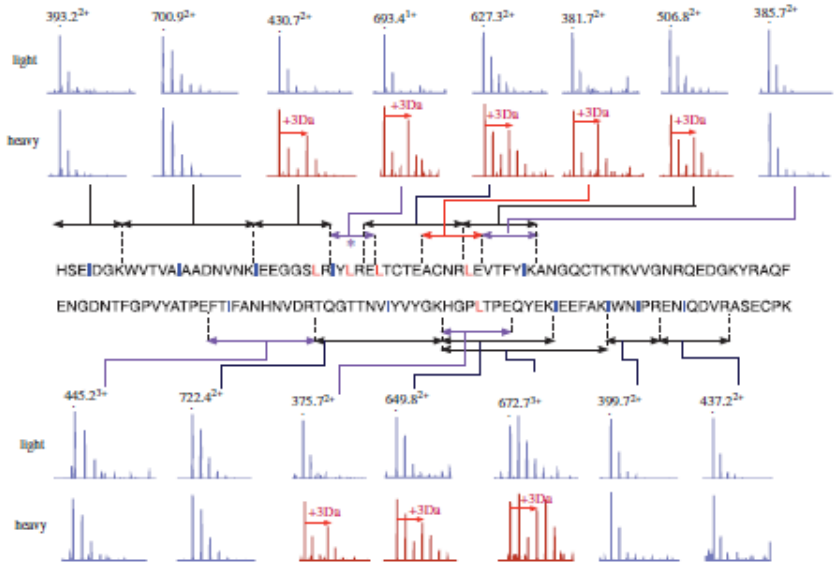


Figure 4. Resolution of leucine and isoleucine by metabolic labelling. After dietary administration of [$^2H_3$] leucine, proteolysis and mass spectrometry of the bank vole lipocalin, the assignment of leucine and isoleucine residues was completed. The figure indicates the residue assignment annotated with the precursor mass spectrum of the appropriate peptide (double-headed arrows), generated from trypsin (black), endopeptidase GluC (red) or digests using both endopeptidases (purple). The monoisotopic, unlabelled ion is marked with a black dot. Spectra that confirm the incorporation of a stable isotope-labelled leucine residue are coloured red and the mass offset (3 Da, due to incorporation of labelled leucine) is indicated with a red arrow.

probasins (prostate expressed 'outlier' lipocalins) and MUPs. A phylogenetic tree (figure 5) defines the relationships between these groups of lipocalins, specifically those from rodents, and a full alignment is given in the electronic supplementary material.

Of interest is the relationship between glareosin and the OBPs (1a, 1b, 2 and 3) that have previously been detected in bank vole urine samples [31]. The four proteins share over 60% sequence identity, and the presence of the lipocalin GXW motif and the disposition of the two disulfide bonds mean that all four proteins share a high level of structural and possibly functional similarity. Yet glareosin was not discovered or described as the predominant urinary protein in the previous study [31], in which the two urinary proteins detected on two-dimensional electrophoresis followed by mass spectrometry were OBP2 and OBP3. Indeed, when we perform a discovery proteomics analysis on a tryptic digest of total urinary proteins, we also see good evidence for these two proteins in bank vole urine (data not shown) but at a much lower level than peptides derived from glareosin. On a one-dimensional SDS–PAGE gel, glareosin is by far the most strongly expressed protein, and at first glance it is not obvious why this protein was not observed in the previous study. However, analysis of the sequence of glareosin and the three OBPs reveals that the predicted isoelectric point (pI) of OBPs 1–3 are 5.0, 4.8 and 4.8, respectively. By contrast, the predicted pI of glareosin is 5.7. In the previous study [31], the pI range of the two-dimensional gel system used to visualize and identify urinary proteins was from 3.9 to 5.1. It is highly likely that glareosin was not resolved by the first, isoelectric focusing dimension, would not have entered the gel and thus could not have been detected.

The complete protein sequence derived by mass spectrometry, including disambiguation of leucine/isoleucine, allowed us to submit the primary sequence to three-dimensional structure prediction. Of predicted structures for glareosin, those produced with a single alignment to aphrodisin [58–60] consistently scored better than those produced with either the top five or top 10 templates identified by HHpred. Aphrodisin is distinctly more closely related to glareosin (47% sequence identity) than other templates (39% at most)—this reinforces the observation that inclusion of more distantly related templates does not always benefit model quality when a closely homologous structure is available. Models generated with the initial HHpred alignment of glareosin with aphrodisin consistently exhibited stereochemical problems near the C-terminus where glareosin has a one-residue deletion compared to aphrodisin. Examination of the aphrodisin structure suggested that side chain interactions would be better retained with a one-residue shift of the deletion position. Positioning the deletion opposite Thr_{149} (mature protein sequence) in the aphrodisin template eliminated serious stereochemical issues and produced better scoring models by validation metrics.

Unexpectedly, the final model set contained two distinct conformations which scored equally well by all criteria. Each conformation gives a normalized QMEAN *Z*-score of 0.44, showing that the structures, by the six distinct component scores considered, perform slightly better than the average protein of a similar size. The two conformations differ in the position of loop 5, connecting β-strands E and F (in the standard family nomenclature [56]). In the 'closed' conformation, the loop lies over the entry to the central binding pocket, as is typically observed in crystal structures (figure 6*a*), while in the 'open' conformation the entrance to the binding pocket is

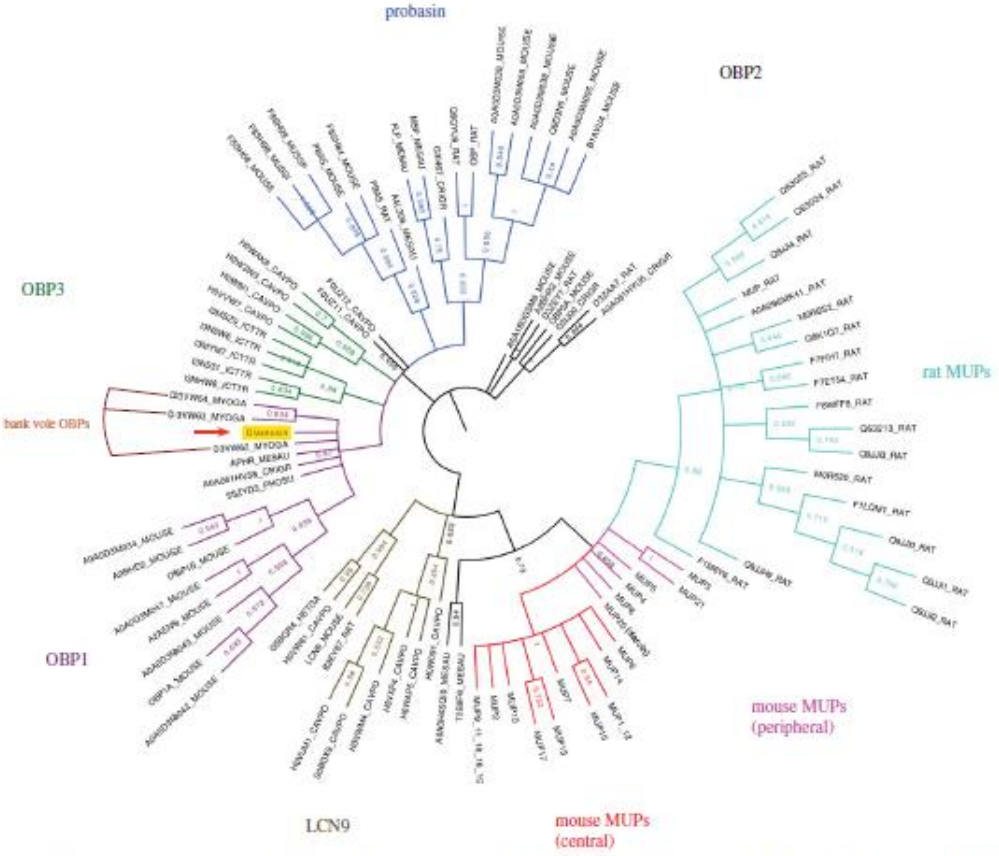


Figure 5. Phylogenetic tree of glareosin-related sequences. Bootstrapped maximum-likelihood phylogenetic tree calculated using MEGA7 as described in Material and methods. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. With the exception of a manually curated set of mouse MUPs based on the MGI database (http://www.informatics.jax.org/searchtool/Search.do?query=mup*), proteins are labelled with UniProt identifiers. The three OBPs previously identified in bank voles [31] are highlighted.

unimpeded and the pocket connects directly with bulk solvent. The validity of the two conformations is supported by the ability of the Rosetta methodology to accurately sample alternative, biologically relevant conformations: it has proved capable of predicting a second allosteric state accurately, given a crystal structure of the first [61]. Pathways of interconversion between these two conformational states could be explored in the future by molecular dynamics simulations. Interestingly, this loop bears a unique one-residue insertion compared to all near relatives of known structure. Thus, it is possible that glareosin has distinct ligand-binding properties when compared to other semiochemical lipocalins whose crystal structures, with cavity occupied or empty, show a strong tendency towards closed structures (figure 6*a*).

The central, β-barrel enclosed cavity of glareosin has a similar volume to aphrodisin; GHECOM [47] estimates them as 305 and 318 Å^3, respectively, while the volumes from Profunc are 357 and 377 Å^3. The cavity of the model structure of glareosin is more elongated, hinting at possible differences in specificity of bound ligands (figure 6*c*). For comparison, GHECOM predicts a cavity of 324 Å^3 and Profunc 410 Å^3 for the unoccupied MUP (1I04.PDB), and GHECOM 396 Å^3 and Profunc 450 Å^3 for a cavity occupied MUP (1I04.PDB). The glareosin cavity is thus of lower volume than the MUP, but is still large enough to accommodate a broad range of low-molecular weight ligands. Ligands of glareosin have yet to be identified.

It has previously been reported that the urinary protein output of *M. glareolus* is sexually dimorphic and that males exhibited obligate proteinuria in all sample types investigated [24]. Males mark new territory in frequent small drops without entirely emptying their bladder, compared with females that deposit large pools of urine [27,28]. This sex-specific behaviour is similar to that of the house mouse, where the repeated marking of territory with small volumes of urine is used to advertise competitive dominance [62,63].

Glareosin appears to be the major protein output in male bank vole urine that is stimulated during the breeding season. As a lipocalin with a clearly defined central cavity that could be switchably accessible, combined with male-specific production and a seasonal expression pattern, this points to a role for glareosin as a major driver of chemical communication between male and female bank voles. As we gain a better

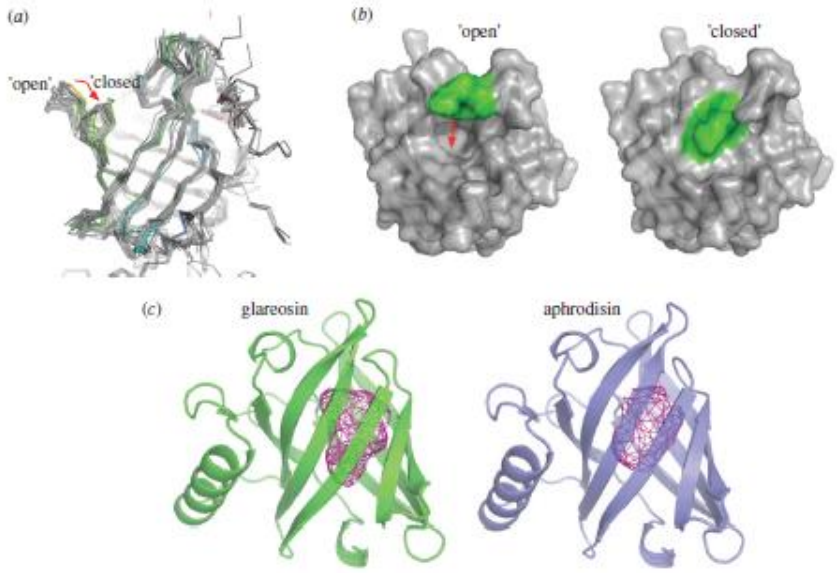


Figure 6. Predicted three-dimensional structure of glareosin. (*a*,*b*) The structure of glareosin was predicted by homology modelling. Two solutions (an 'open' and a 'closed' conformation) were predicted equally well. In (*a*), the two solutions are coloured blue to red from N- to C-terminus with all experimental structures of lipocalins sharing at least 25% sequence identity with glareosin shown in grey. In (*b*), the loop differing in conformation is shown as green, and the rest of the glareosin models as grey. (*c*) The cavity at the centre of the closed glareosin structure was analysed using the Profunc server [48] and compared with aphrodisin.

understanding of the use of lipocalins in chemical communication in rodents, an interesting bifurcation is increasingly evident. Of rodents, Muridae (Old World mice, rats) have evolved polymorphic families of MUPs that create considerable potential for individual variation in proteins—they function as pheromone-binding proteins but also as pheromones in their own right. Currently, our knowledge is largely derived from studies of house mice (*M. musculus*) and brown rats (*Rattus norvegicus*). By contrast, Cricetinae (hamsters, voles) also elaborate protein in their secretions, but evidence thus far suggests that this is restricted to high levels of a single protein. Thus, roborovskin, from *Phodopus roborovskii*, is a single lipocalin produced in the urine equally by both sexes [64]. The vaginal discharge of the golden hamster, *Mesocricetus auratus*, contains abundant levels of the lipocalin aphrodisin, which acts as a pheromone (possibly in concert with a bound ligand) to stimulate copulatory behaviour by males [8,57,58,60,65]. Aphrodisin is a female-specific lipocalin in vaginal secretions, whereas glareosin is a male-specific protein restricted to the breeding season. While none of these species invoke the same polymorphic variation as MUPs as the Muridae, it is probable that clear functions in intraspecific communication will be found. Interestingly, Bathyerginae (*Fukomys*, naked mole rat) also seem to express urinary proteins that are more aphrodisin-like [66]. It is possible that MUP-like sequences have evolved different roles to aphrodisin/OBP-like proteins, and that in muroid rodents, a high level of polymorphism may be a unique feature. Whereas MUPs are readily identified and classified within the lipocalin family, there is a need for clearer understanding of the aphrodisin-like proteins. OBPs are expressed in nasal tissue in a wide range of species [67–72] and may facilitate the transport of low-molecular-weight signalling molecules across the mucosal membrane. However, OBPs are now being increasingly reported in the urine of rodents, and it is likely that they are also involved in the generation as well as the reception of chemosignals. Further study of the role of lipocalins in chemical communication seems likely to reveal a breadth of mechanisms whereby information is conveyed between conspecifics.

Ethics. All procedures involved in this study were non-invasive. Animal trapping, use and care were in accordance with EU directive 2010/63/EU, UK Home Office code of practice for the housing and care of animals bred, supplied or used for scientific purposes, and UK research funder's guidelines on responsibility in the use of animals in bioscience research. The University of Liverpool Animal Welfare Committee approved the work, but no specific licences were required.

Data accessibility. Detailed methods are presented in the electronic supplementary material. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [32] partner repository with the dataset identifier PXD006645 and 10.6019/PXD006645.

Authors' contributions. G.M.L., J.U., K.S.L., A.J., D.M.S. and M.J.T. carried out the experimental work and participated in data analysis. D.J.R. conducted all molecular modelling. R.J.B. and J.L.H. conceived the study, designed the study and coordinated the study. All authors helped draft the manuscript and all authors gave final approval for publication.

Competing interests. We declared that we have no competing interest.

Funding. This work was funded in part by BBSRC (BB/J002631/1 and BB/M012557/1).

Acknowledgements. We are grateful to Dr Richard Humphries, Amanda Davidson and the animal care staff at the Mammalian Behaviour and Evolution Laboratory, and Dr Philip Brownridge at the Centre for Proteome Research for excellent instrument care. J.U. is grateful to the Biotechnology and Biological Sciences Research Council (BBSRC) for provision of a PhD studentship.

References

1. Brown RE, Macdonald DW. 1985 *Social odours in mammals*. Oxford, UK: Clarendon Press.
2. Gosling LM, Roberts SC. 2001 Scent-marking by male mammals: cheat-proof signals to competitors and mates. *Adv. Study Behav.* **30**, 169–217. (doi:10.1016/S0065-3454(01)80007-3)
3. Hurst JL. 1993 The priming effects of urine substrate marks on interactions between male house mice, *Mus musculus domesticus* Schwarz & Schwarz. *Anim. Behav.* **45**, 55–61. (doi:10.1006/anbe.1993.1007)
4. Hurst JL, Beynon RJ, Humphries RE, Malone N, Nevison CM, Payne CE, Robertson DHL, Veggerby C. 2001 Information in scent signals of competitive social status: the interface between behaviour and chemistry. *Chem. Signals Vertebr.* **9**, 43–52. (doi:10.1007/978-1-4615-0671-3_6)
5. Petrulis A. 2013 Chemosignals, hormones and mammalian reproduction. *Horm. Behav.* **68**, 723–741. (doi:10.1016/j.yhbeh.2013.03.011)
6. Hurst JL, Beynon RJ. 2013 Rodent urinary proteins: genetic identity signals and pheromones. *Chem. Signals Vertebr.* **12**, 117–133. (doi:10.1007/978-1-4614-5927-9_9)
7. Lazar J, Greenwood DR, Rasmussen LE, Prestwich GD. 2002 Molecular and functional characterization of an odorant binding protein of the Asian elephant, *Elephas maximus*: implications for the role of lipocalins in mammalian olfaction. *Biochemistry* **41**, 11786–11794. (doi:10.1021/bi025673+)
8. Singer AG, Macides F, Clancy AN, Agosta WC. 1986 Purification and analysis of a proteinaceous aphrodisiac pheromone from hamster vaginal discharge. *J. Biol. Chem.* **261**, 13323–13326.
9. Stopka P, Kuntová B, Klempt P, Havrdová L, Černá M, Stopková R. 2016 On the saliva proteome of the Eastern European house mouse (*Mus musculus musculus*) focusing on sexual signalling and immunity. *Sci. Rep.* **6**, 32481. (doi:10.1038/srep32481)
10. Beynon RJ, Veggerby C, Payne CE, Robertson DH, Gaskell SJ, Humphries RE, Hurst JL. 2002 Polymorphism in major urinary proteins: molecular heterogeneity in a wild mouse population. *J. Chem. Ecol.* **28**, 1429–1446. (doi:10.1023/A:1016252703836)
11. Chamero P, Marton TF, Logan DW, Flanagan K, Cruz JR, Saghatelian A, Cravatt BF, Stowers L. 2007 Identification of protein pheromones that promote aggressive behaviour. *Nature* **450**, 899–902. (doi:10.1038/nature05997)
12. Cheetham SA, Thom MD, Jury F, Ollier WE, Beynon RJ, Hurst JL. 2007 The genetic basis of individual-recognition signals in the mouse. *Curr. Biol.* **17**, 1771–1777. (doi:10.1016/j.cub.2007.10.007)
13. Garratt M, Stockley P, Armstrong SD, Beynon RJ, Hurst JL. 2011 The scent of senescence: sexual signalling and female preference in house mice. *J. Evol. Biol.* **24**, 2398–2409. (doi:10.1111/j.1420-9101.2011.02367.x)
14. Hurst JL, Payne CE, Nevison CM, Marie AD, Humphries RE, Robertson DH, Cavaggioni A, Beynon RJ. 2001 Individual recognition in mice mediated by major urinary proteins. *Nature* **414**, 631–634. (doi:10.1038/414631a)
15. Lopes PC, Könıg B. 2016 Choosing a healthy mate: sexually attractive traits as reliable indicators of current disease status in house mice. *Anim. Behav.* **111**, 119–126. (doi:10.1016/j.anbehav.2015.10.011)
16. Nelson AC, Cunningham CB, Ruff JS, Potts WK. 2015 Protein pheromone expression levels predict and respond to the formation of social dominance networks. *J. Evol. Biol.* **28**, 1213–1224. (doi:10.1111/jeb.12643)
17. Sheehan MJ, Lee V, Corbett-Detig R, Bi K, Beynon RJ, Hurst JL, Nachman MW. 2016 Selection on coding and regulatory variation maintains individuality in major urinary protein scent marks in wild mice. *PLoS Genet.* **12**, e1005891. (doi:10.1371/journal.pgen.1005891)
18. Stopka P, Janotova K, Heyrovsky D. 2007 The advertisement role of major urinary proteins in mice. *Physiol. Behav.* **91**, 667–670. (doi:10.1016/j.physbeh.2007.03.030)
19. Thom MD, Stockley P, Jury F, Ollier WE, Beynon RJ, Hurst JL. 2008 The direct assessment of genetic heterozygosity through scent in the mouse. *Curr. Biol.* **18**, 619–623. (doi:10.1016/j.cub.2008.03.056)
20. Beynon RJ, Hurst JL. 2004 Urinary proteins and the modulation of chemical scents in mice and rats. *Peptides* **25**, 1553–1563. (doi:10.1016/j.peptides.2003.12.025)
21. Hurst JL, Robertson DHL, Tolladay U, Beynon RJ. 1998 Proteins in urine scent marks of male house mice extend the longevity of olfactory signals. *Anim. Behav.* **55**, 1289–1297. (doi:10.1006/anbe.1997.0650)
22. Gómez-Baena G, Armstrong SD, Phelan MM, Hurst JL, Beynon RJ. 2014 The major urinary protein system in the rat. *Biochem. Soc. Trans.* **42**, 886–892. (doi:10.1042/BST20140083)
23. Vandoren G, Mertens B, Heyns W, Van Baelen H, Rombauts W, Verhoeven G. 1983 Different forms of alpha 2u-globulin in male and female rat urine. *Eur. J. Biochem.* **134**, 175–181. (doi:10.1111/j.1432-1033.1983.tb07548.x)
24. Kruczek M, Marchlewska-Koj A. 1985 Androgen-dependent proteins in the urine of bank voles (*Clethrionomys glareolus*). *J. Reprod. Fertil.* **75**, 189–192. (doi:10.1530/jrf.0.0750189)
25. Bujalska G. 1990 Social system of the bank vole, *Clethrionomys glareolus*. In *Social systems and population cycles in voles* (eds RHD Tamarin, RS Ostfeld, SR Pugh, G Bujalska), pp. 155–167. Berlin, Germany: Springer.
26. Mazurkiewicz M. 1981 Spatial organization of a bank vole population in years of small or large numbers. *Acta Theriol.* **26**, 31–45. (doi:10.4098/AT.arch.81-3)
27. Christiansen E. 1980 Urinary marking in wild bank voles, *Clethrionomys glareolus* in relation to season and sexual status. *Behav. Neural Biol.* **28**, 123–127. (doi:10.1016/S0163-1047(80)93258-6)
28. Johnson RP. 1975 Scent marking with urine in two races of the bank vole (*Clethrionomys glareolus*). *Behaviour* **55**, 81–98. (doi:10.1163/156853975X00425)
29. Verplanckе G, LeBoulenge E. 2011 Differential marking, investigation and motor activity in presence of conspecific odours differing on their population of origin in bank voles. *Environ. Sci.* **1**, 57–69.
30. Kruczek M. 1997 Male rank and female choice in the bank vole, *Clethrionomys glareolus*. *Behav. Processes* **40**, 171–176. (doi:10.1016/S0376-6357(97)00785-7)
31. Stopková R, Zdráhal Z, Ryba S, Sedo O, Sandera M, Stopka P. 2010 Novel OBP genes similar to hamster Aphrodisin in the bank vole, *Myodes glareolus*. *BMC Genomics* **11**, 45. (doi:10.1186/1471-2164-11-45)
32. Vizcaíno JA *et al.* 2016 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, 11033. (doi:10.1093/nar/gkw880)
33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)
34. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786. (doi:10.1038/nmeth.1701)
35. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)
36. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009 Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191. (doi:10.1093/bioinformatics/btp033)
37. Jones DT, Taylor WR, Thornton JM. 1992 The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282. (doi:10.1093/bioinformatics/8.3.275)
38. Felsenstein J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791. (doi:10.1111/j.1558-5646.1985.tb00420.x)
39. Kumar S, Stecher G, Tamura K. 2016 MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874. (doi:10.1093/molbev/msw054)
40. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D. 2013 High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742. (doi:10.1016/j.str.2013.08.005)

41. Söding J, Biegert A, Lupas AN. 2005 The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248. (doi:10.1093/nar/gki408)
42. Sippl MJ. 1993 Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362. (doi:10.1002/prot.340170404)
43. Shen MY, Sali A. 2006 Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524. (doi:10.1110/ps.062416606)
44. Benkert P, Künzli M, Schwede T. 2009 QMEAN server for protein model quality estimation. *Nucleic Acids Res.* **37**, W510–W514. (doi:10.1093/nar/gkp322)
45. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. 1993 PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291. (doi:10.1107/S0021889892009944)
46. Krissinel E. 2012 Enhanced fold recognition using efficient short fragment clustering. *J. Mol. Biochem.* **1**, 76–85.
47. Kawabata T. 2010 Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **78**, 1195–1211. (doi:10.1002/prot.22639)
48. Laskowski RA, Watson JD, Thornton JM. 2005 ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33**, W89–W93. (doi:10.1093/nar/gki414)
49. Evershed RP, Robertson DH, Beynon RJ, Green BN. 1993 Application of electrospray ionization mass spectrometry with maximum-entropy analysis to allelic 'fingerprinting' of major urinary proteins. *Rapid Commun. Mass Spectrom.* **7**, 882–886. (doi:10.1002/rcm.1290071005)
50. Robertson DH, Cox KA, Gaskell SJ, Evershed RP, Beynon RJ. 1996 Molecular heterogeneity in the major urinary proteins of the house mouse *Mus musculus*. *Biochem. J.* **316**, 265–272. (doi:10.1042/bj3160265)
51. Mudge JM, Armstrong SD, McLaren K, Beynon RJ, Hurst JL, Nicholson C, Robertson DH, Wilming LG, Harrow JL. 2008 Dynamic instability of the major urinary protein gene family revealed by genomic and phenotypic comparisons between C57 and 129 strain mice. *Genome Biol.* **9**, R91. (doi:10.1186/gb-2008-9-5-r91)
52. Beynon RJ *et al.* 2015 Mass spectrometry for structural analysis and quantification of the major urinary proteins of the house mouse. *Int. J. Mass Spectrom.* **391**, 146–156. (doi:10.1016/j.ijms.2015.07.026)
53. Robertson DH, Hurst JL, Bolgar MS, Gaskell SJ, Beynon RJ. 1997 Molecular heterogeneity of urinary proteins in wild house mouse populations. *Rapid Commun. Mass Spectrom.* **11**, 786–790. (doi:10.1002/(SICI)1097-0231(19970422)11:7<786:AID-RCM876>3.0.CO;2-8)
54. Hurst JL, Beynon RJ, Armstrong SD, Davidson AJ, Roberts SA, Gómez-Baena G, Smadja CM, Ganem G. 2017 Molecular heterogeneity in major urinary proteins of *Mus musculus* subspecies: potential candidates involved in speciation. *Sci. Rep.* **7**, 44992. (doi:10.1038/srep44992)
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1016/S0022-2836(05)80360-2)
56. Flower DR, North AC, Sansom CE. 2000 The lipocalin protein family: structural and sequence overview. *Biochim. Biophys. Acta* **1482**, 9–24. (doi:10.1016/S0167-4838(00)00148-5)
57. Vincent F, Löbel D, Brown K, Spinelli S, Grote P, Breer H, Cambillau C, Tegoni M. 2001 Crystal structure of aphrodisin, a sex pheromone from female hamster. *J. Mol. Biol.* **305**, 459–469. (doi:10.1006/jmbi.2000.4241)
58. Henzel WJ, Rodriguez H, Singer AG, Stults JT, Macides F, Agosta WC, Niall H. 1988 The primary structure of aphrodisin. *J. Biol. Chem.* **263**, 16 682–16 687.
59. Mägert HJ, Gieslak A, Alkan O, Lüscher B, Kauffels W, Forssmann WG. 1999 The golden hamster aphrodisin gene. Structure, expression in parotid glands of female animals, and comparison with a similar murine gene. *J. Biol. Chem.* **274**, 444–450. (doi:10.1074/jbc.274.1.444)
60. Briand L, Huet J, Perez V, Lenoir G, Nespoulous C, Boucher Y, Trotier D, Pernollet JC. 2000 Odorant and pheromone binding by aphrodisin, a hamster aphrodisiac protein. *FEBS Lett.* **476**, 179–185. (doi:10.1016/S0014-5793(00)01719-1)
61. Kidd BA, Baker D, Thomas WE. 2009 Computation of conformational coupling in allosteric proteins. *PLoS Comput. Biol.* **5**, e1000484. (doi:10.1371/journal.pcbi.1000484)
62. Desjardins C, Maruniak JA, Bronson FH. 1973 Social rank in house mice: differentiation revealed by ultraviolet visualization of urinary marking patterns. *Science* **182**, 939–941. (doi:10.1126/science.182.4115.939)
63. Hurst JL, Beynon RJ. 2004 Scent wars: the chemobiology of competitive signalling in mice. *Bioessays* **26**, 1288–1298. (doi:10.1002/bies.20147)
64. Turton MJ, Robertson DH, Smith JR, Hurst JL, Beynon RJ. 2010 Roborovskin, a lipocalin in the urine of the Roborovski hamster, *Phodopus roborovskii*. *Chem. Senses* **35**, 675–684. (doi:10.1093/chemse/bjq060)
65. Briand L, Trotier D, Pernollet JC. 2004 Aphrodisin, an aphrodisiac lipocalin secreted in hamster vaginal secretions. *Peptides* **25**, 1545–1552. (doi:10.1016/j.peptides.2003.10.026)
66. Hagemeyer P, Begall S, Janotova K, Todrank J, Heth G, Jedelsky PL, Burda H, Stopka P. 2011 Searching for major urinary proteins (MUPs) as chemosignals in urine of subterranean rodents. *J. Chem. Ecol.* **37**, 687–694. (doi:10.1007/s10886-011-9971-y)
67. Briand L, Nespoulous C, Perez V, Rémy JJ, Huet JC, Pernollet JC. 2000 Ligand-binding properties and structural characterization of a novel rat odorant-binding protein variant. *Eur. J. Biochem.* **267**, 3079–3089. (doi:10.1046/j.1432-1033.2000.01340.x)
68. Garibotti M, Navarrini A, Pisanelli AM, Pelosi P. 1997 Three odorant-binding proteins from rabbit nasal mucosa. *Chem. Senses* **22**, 383–390. (doi:10.1093/chemse/22.4.383)
69. Heydel JM *et al.* 2013 Odorant-binding proteins and xenobiotic metabolizing enzymes: implications in olfactory perireceptor events. *Anat. Rec.* **296**, 1333–1345. (doi:10.1002/ar.22735)
70. Löbel D, Strotmann J, Jacob M, Breer H. 2001 Identification of a third rat odorant-binding protein (OBP3). *Chem. Senses* **26**, 673–680. (doi:10.1093/chemse/26.6.673)
71. Pes D, Pelosi P. 1995 Odorant-binding proteins of the mouse. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **112**, 471–479. (doi:10.1016/0305-0491(95)00063-1)
72. Tegoni M, Pelosi P, Vincent F, Spinelli S, Campanacci V, Grolli S, Ramoni R, Cambillau C. 2000 Mammalian odorant binding proteins. *Biochim. Biophys. Acta* **1482**, 229–240. (doi:10.1016/S0167-4838(00)00167-9)

3.6.2 Individual Intact Mass Spectra

3.6.2.1 Captive bank vole urine

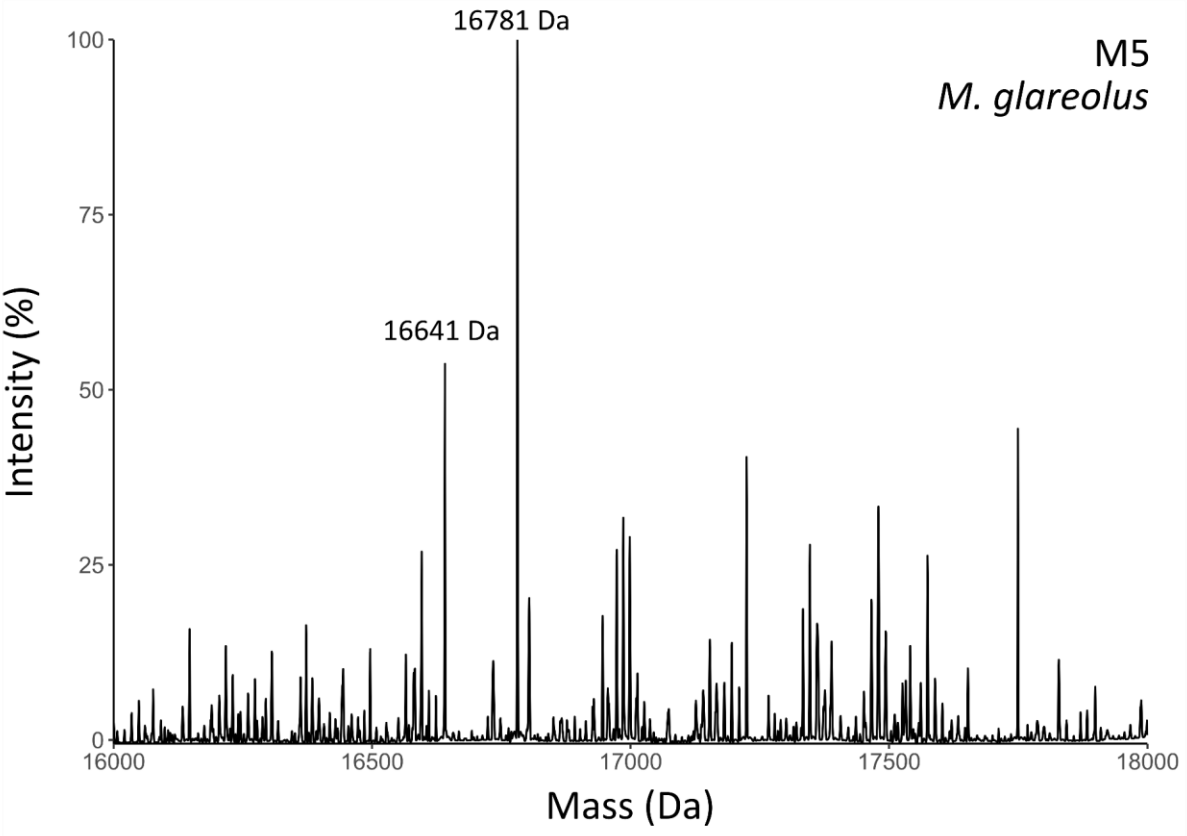


Figure 3.1 Intact mass analysis of urine from captive male bank vole 5.

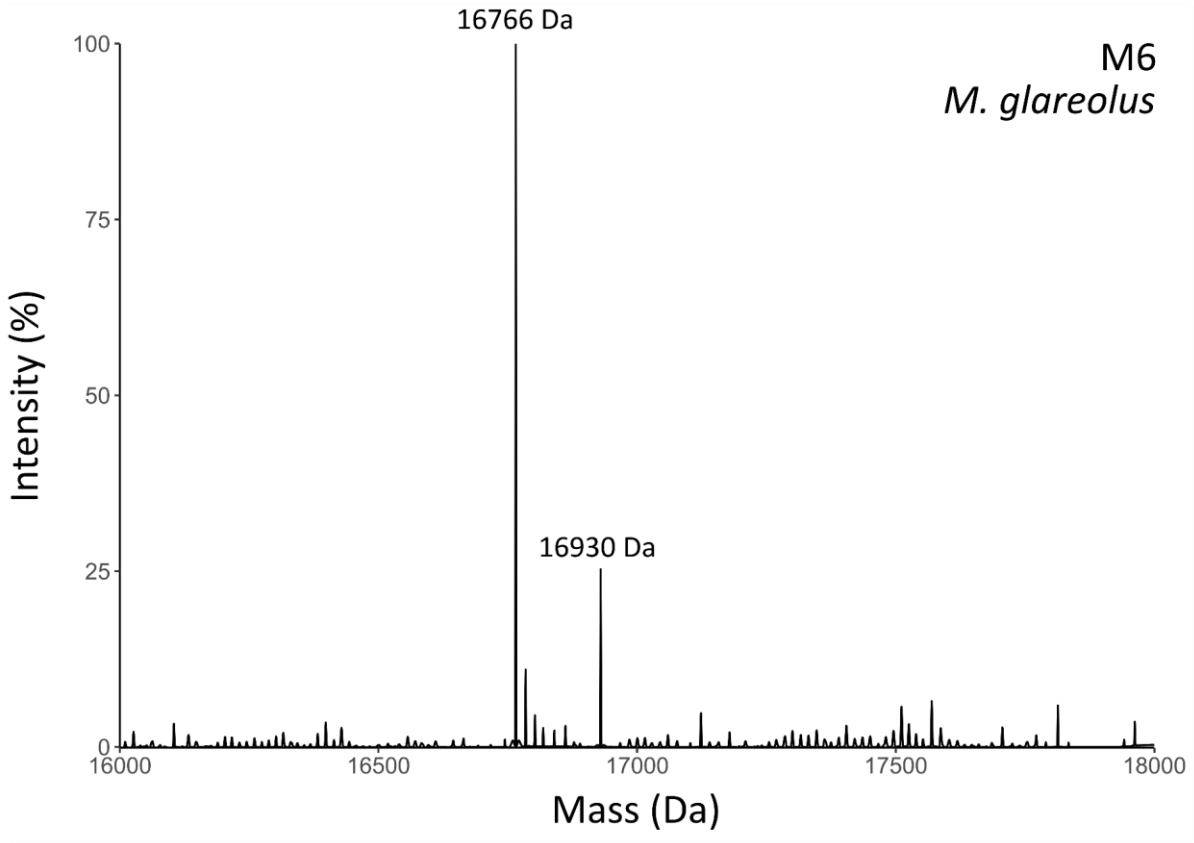


Figure 3.2 Intact mass analysis of urine from captive male bank vole 6.

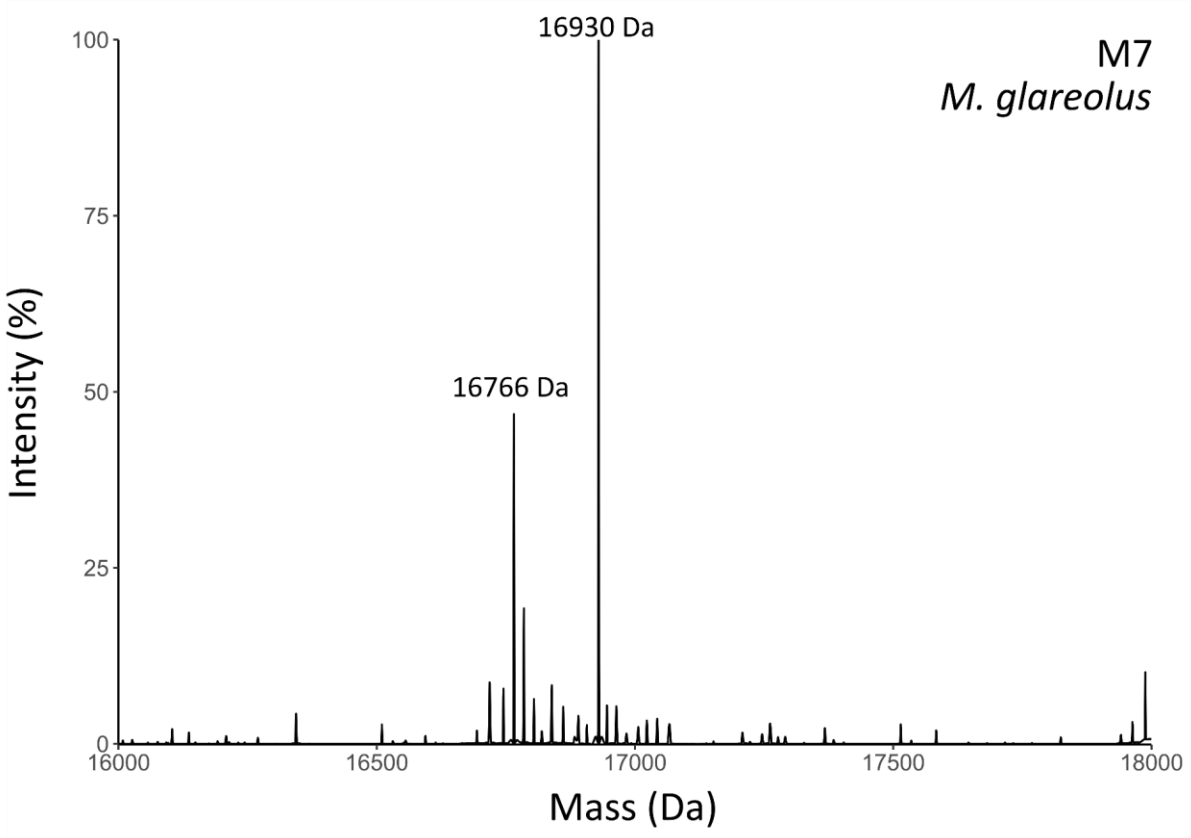


Figure 3.3 Intact mass analysis of urine from captive male bank vole 7.

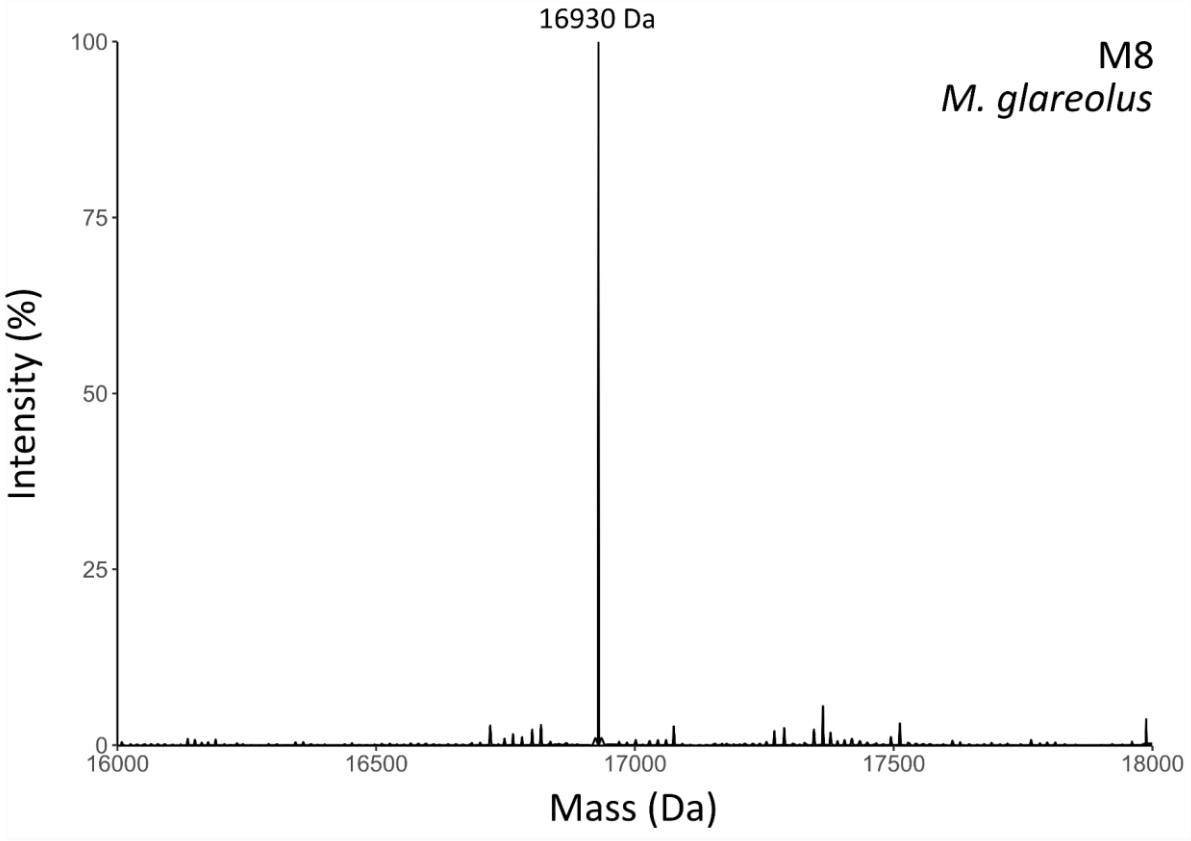


Figure 3.4 Intact mass analysis of urine from captive male bank vole 8.

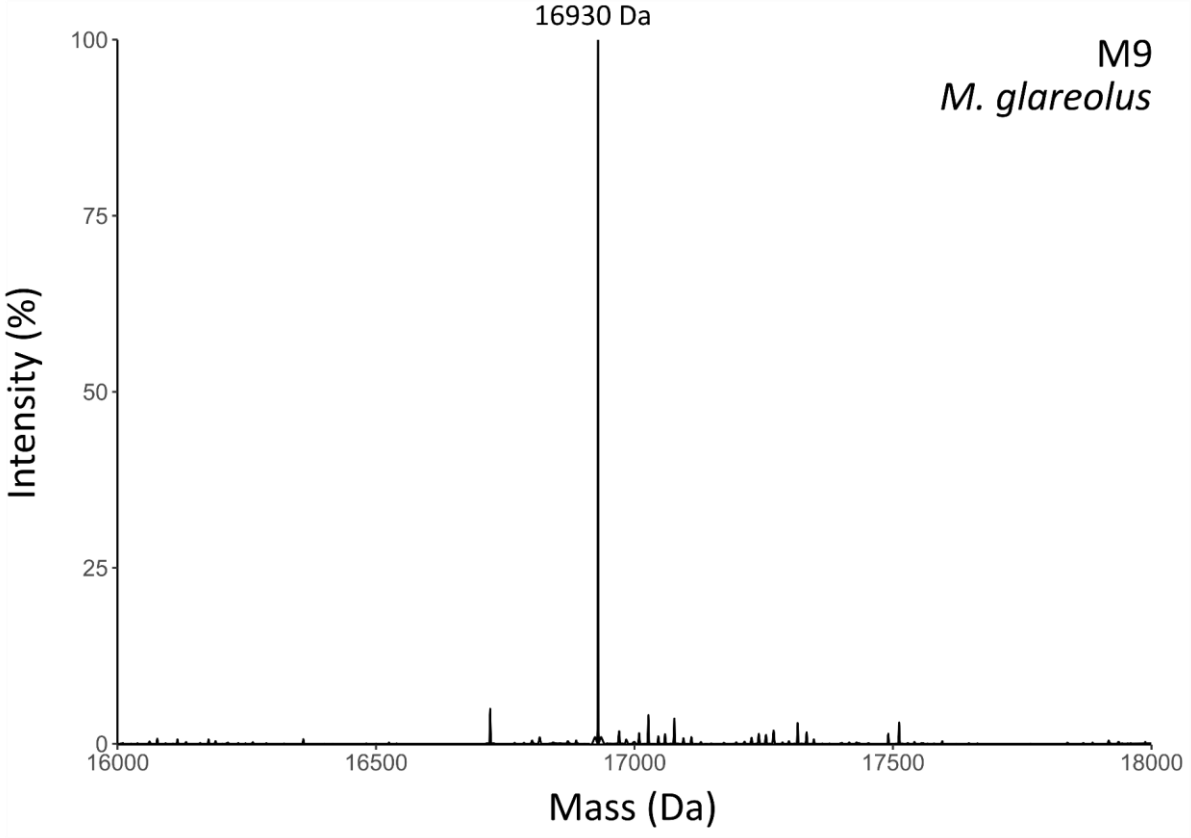


Figure 3.5 Intact mass analysis of urine from captive male bank vole 9.

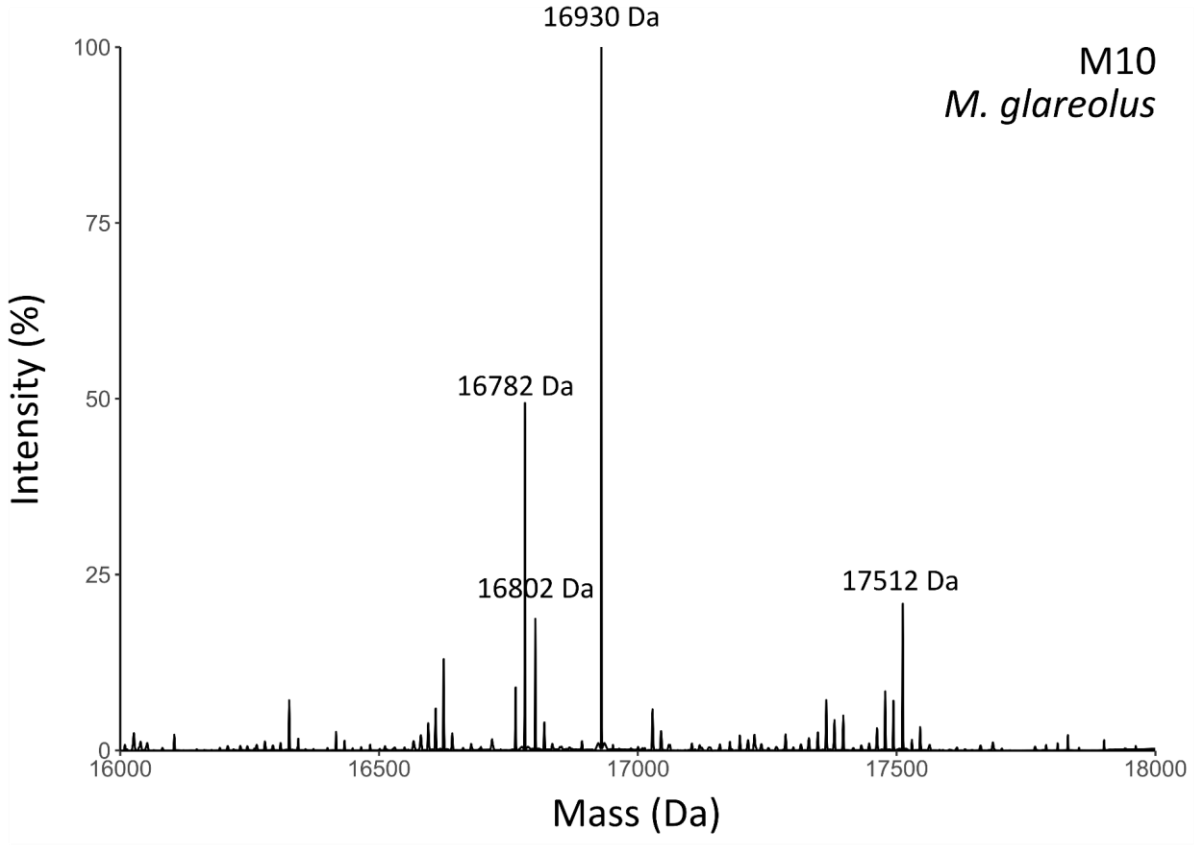


Figure 3.6 Intact mass analysis of urine from captive male bank vole 10.

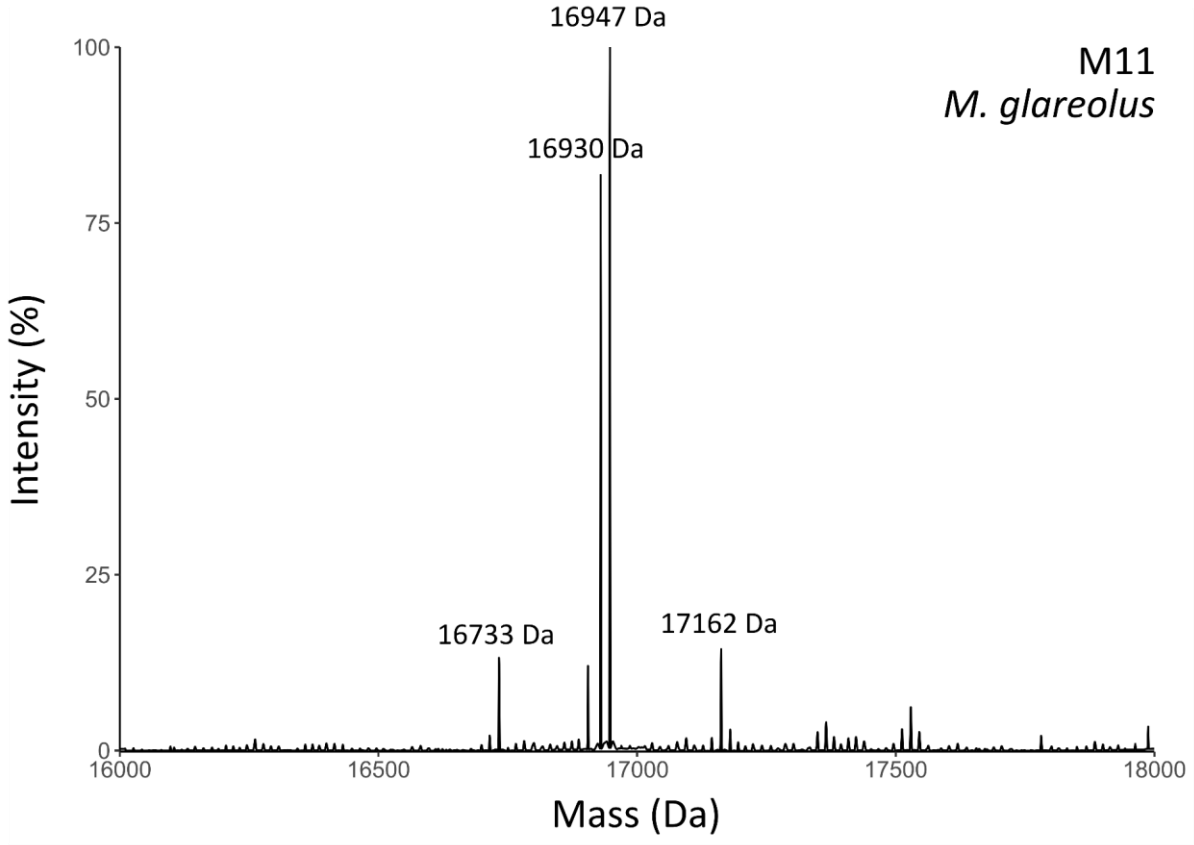


Figure 3.7 Intact mass analysis of urine from captive male bank vole 11.

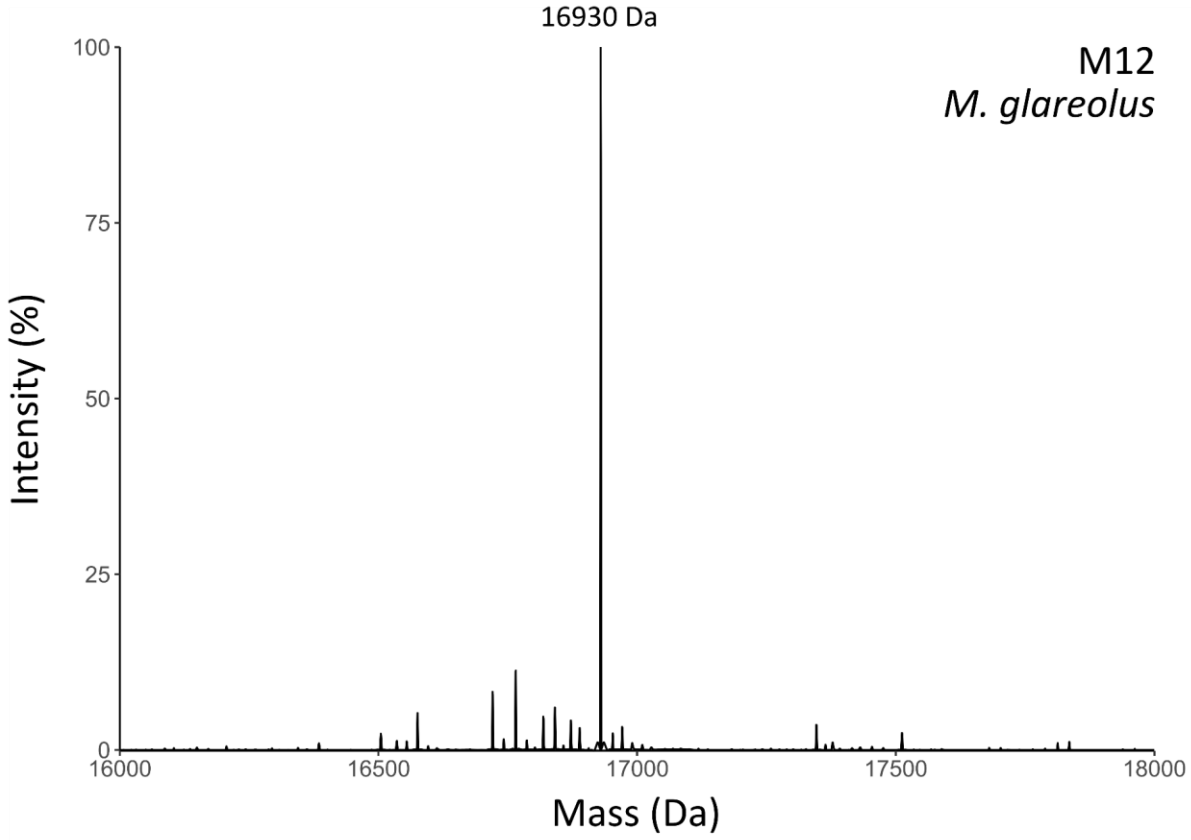


Figure 3.8 Intact mass analysis of urine from captive male bank vole 12.

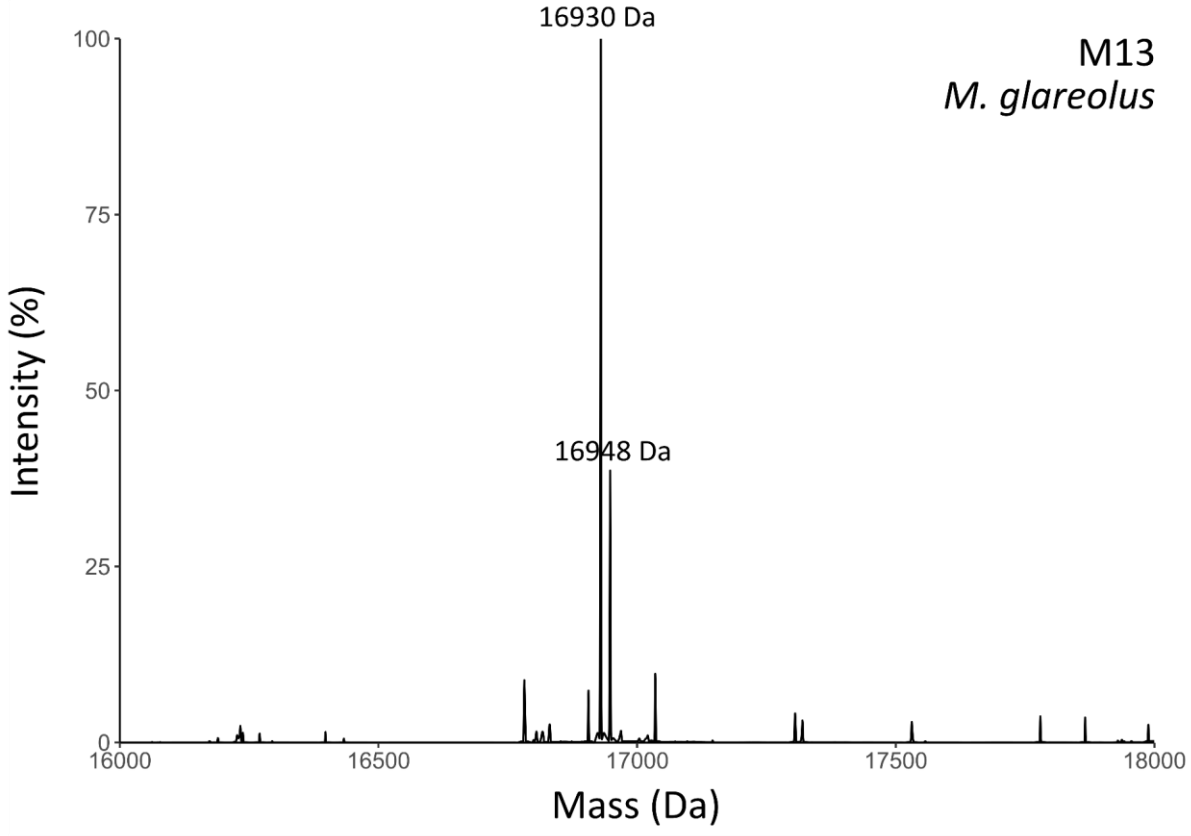


Figure 3.9 Intact mass analysis of urine from captive male bank vole 13

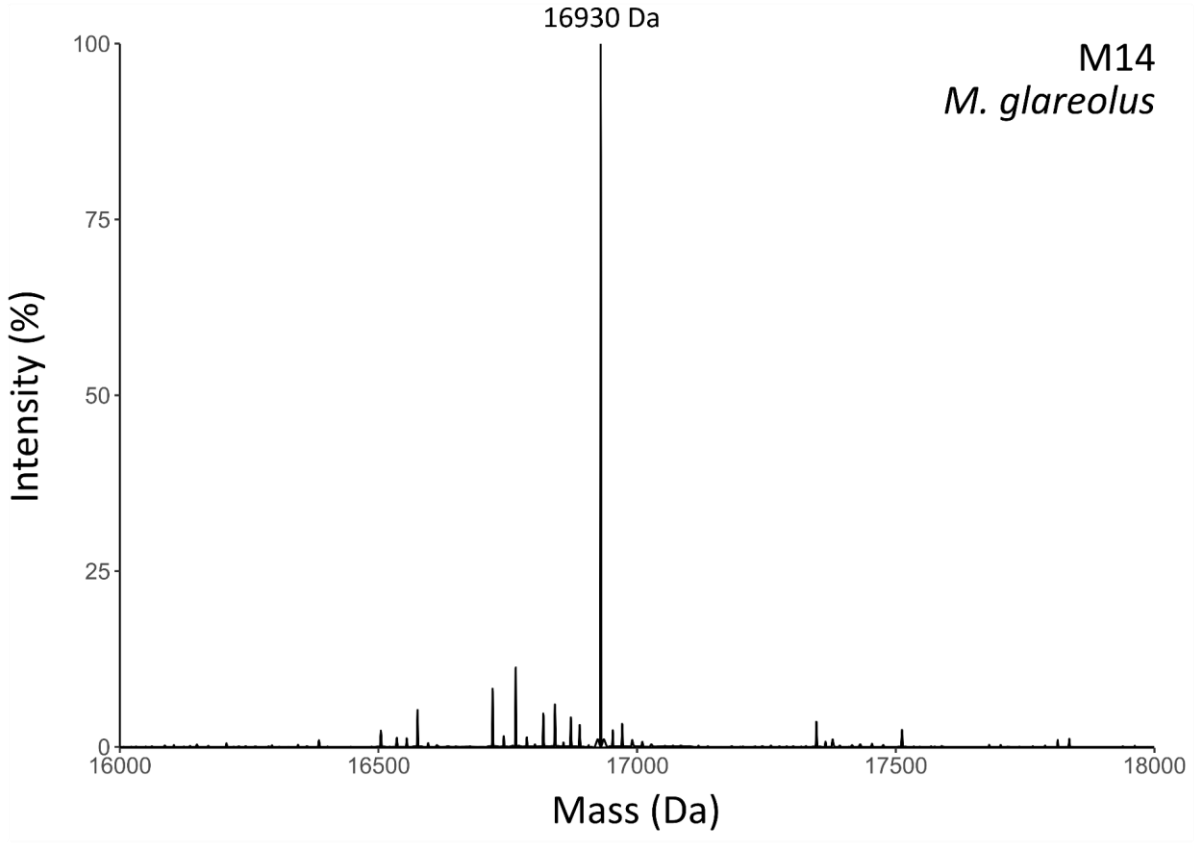


Figure 3.10 Intact mass analysis of urine from captive male bank vole 14.

3.6.2.2 *Wild bank vole urine*

Labelling of the following intact mass spectra correspond to that of the SDS-PAGE image below (also seen in Figure 3.3).

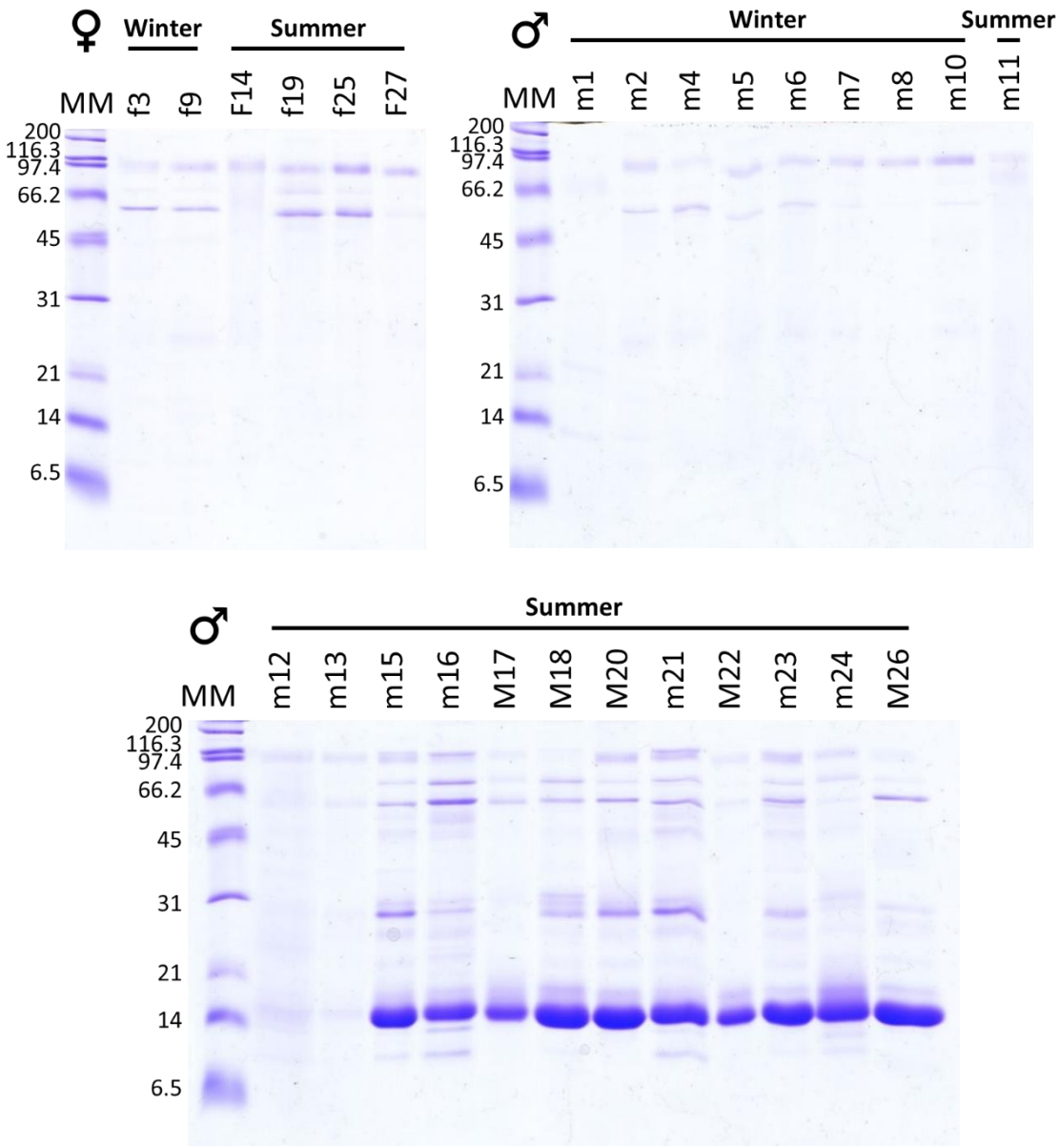


Figure 3.11 SDS-PAGE analysis of urine from wild mature and immature male bank voles.

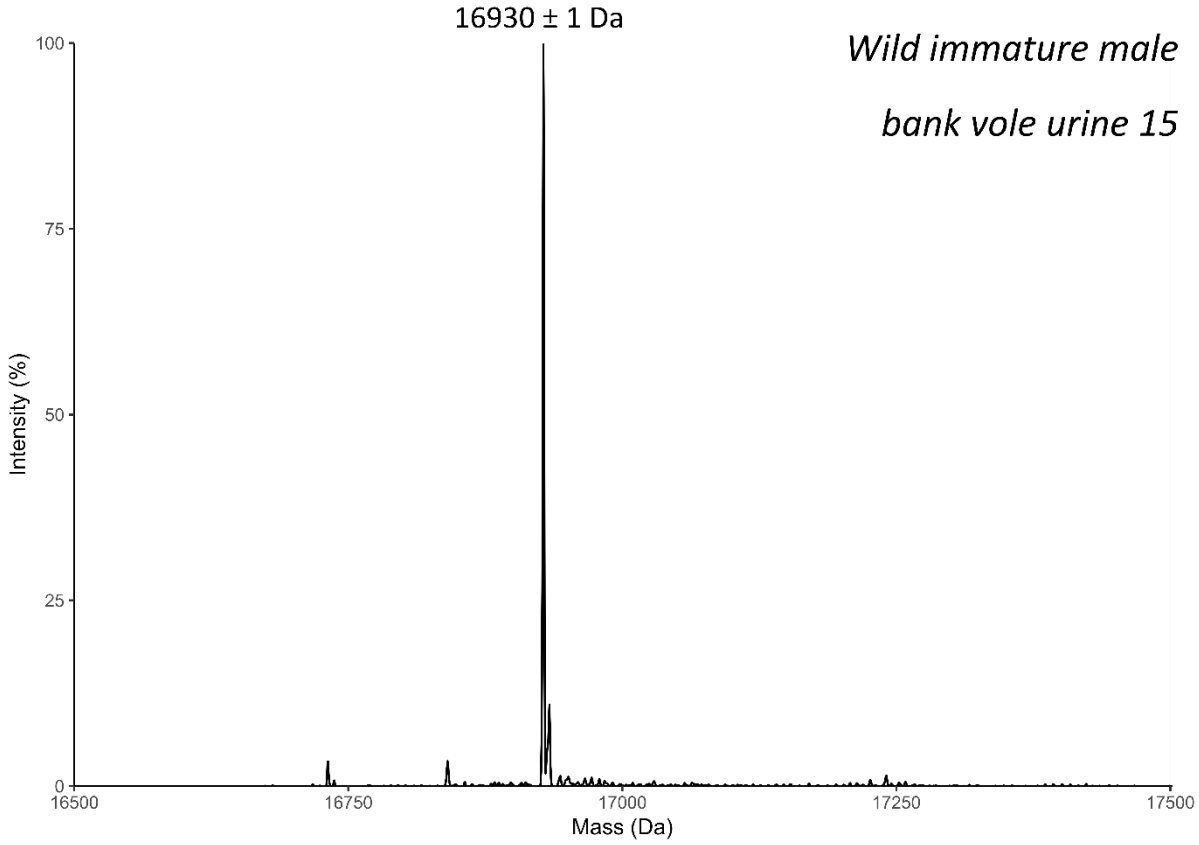


Figure 3.12 Intact mass analysis of urine from wild immature male bank vole 15.

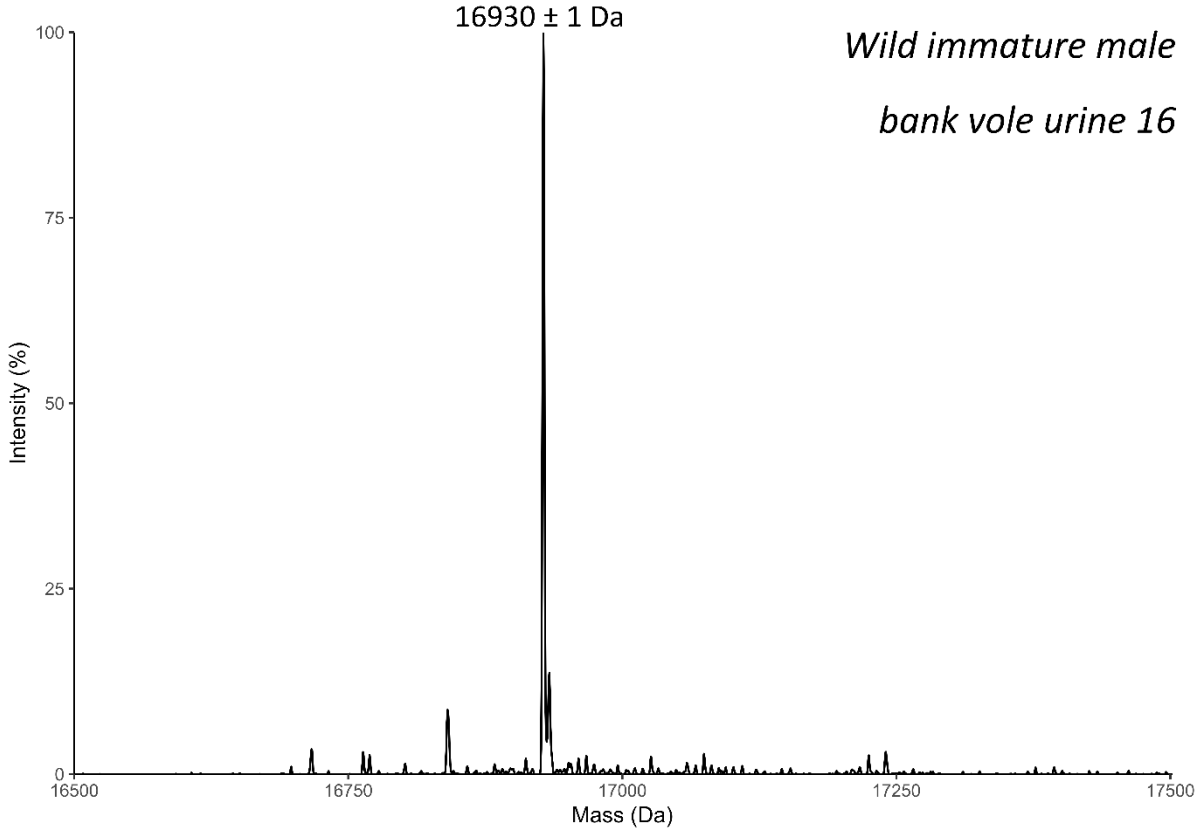


Figure 3.13 Intact mass analysis of urine from wild immature male bank vole 16.

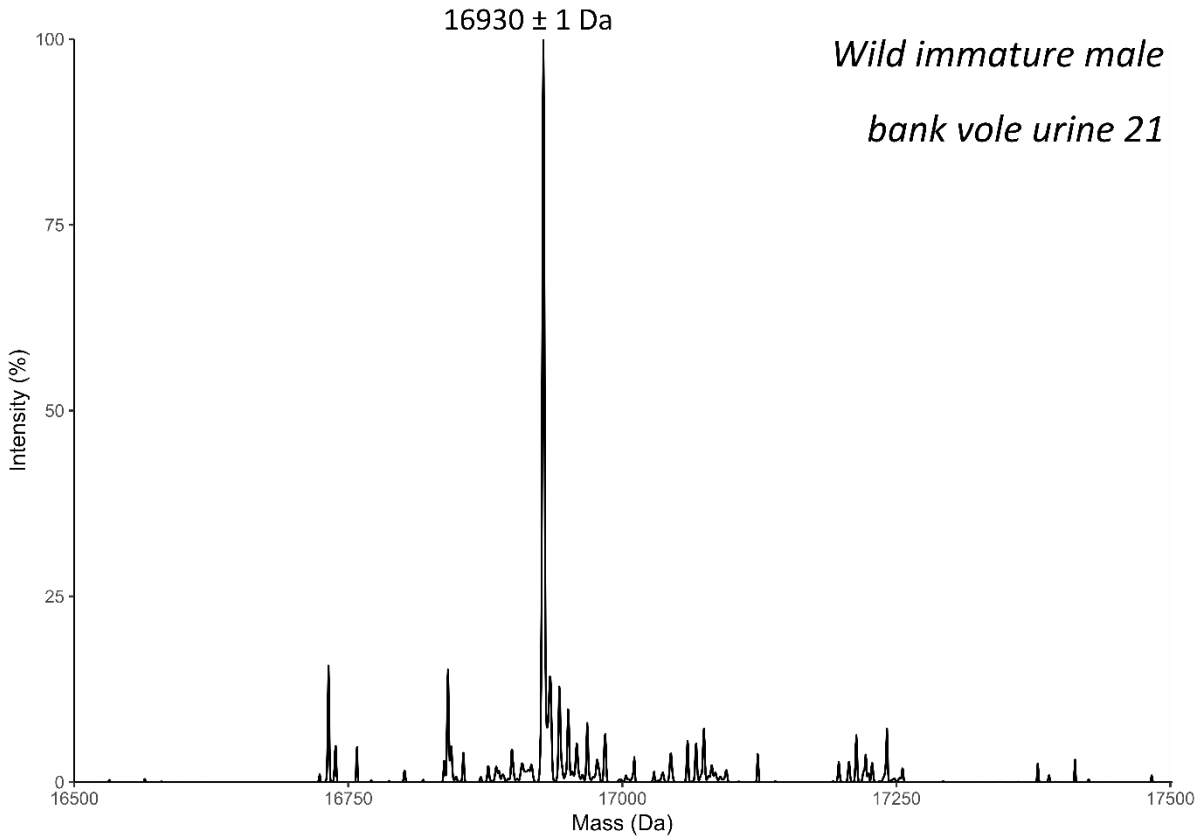


Figure 3.14 Intact mass analysis of urine from wild immature male bank vole 21.

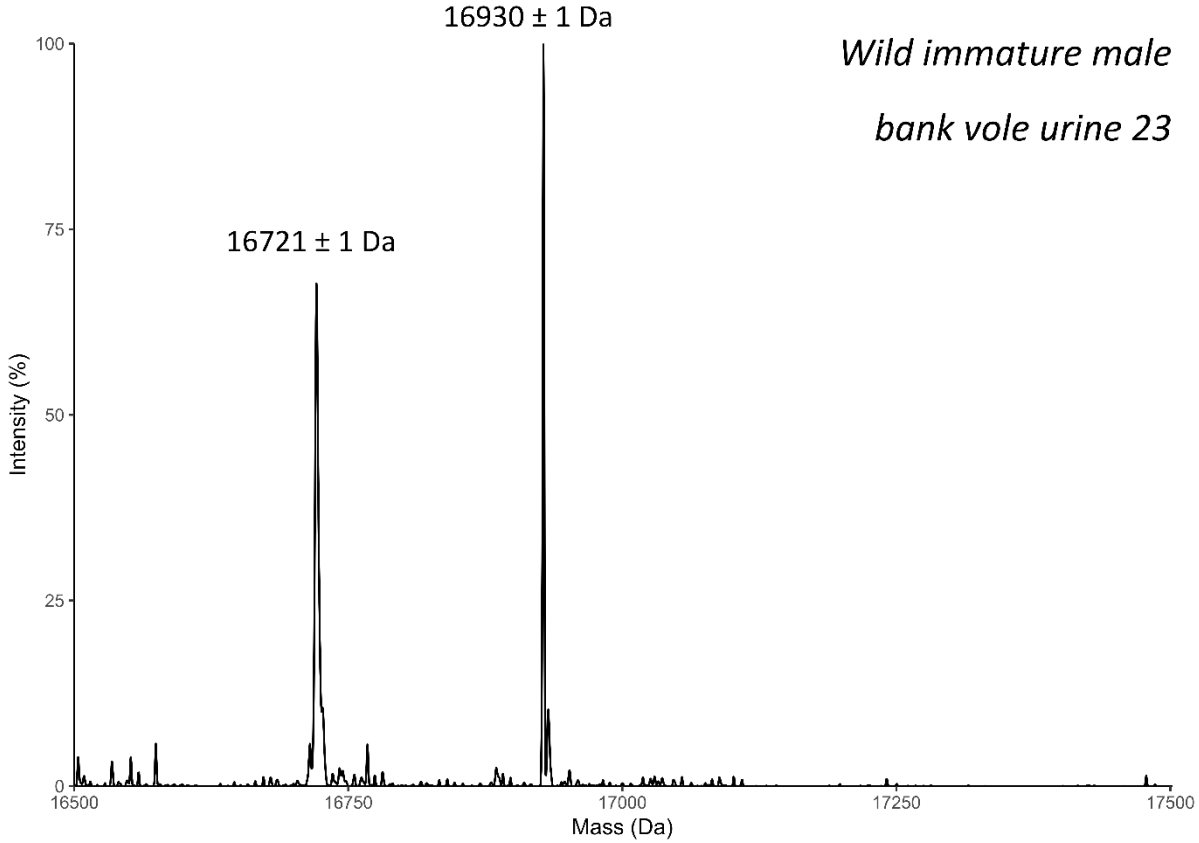


Figure 3.15 Intact mass analysis of urine from wild immature male bank vole 23.

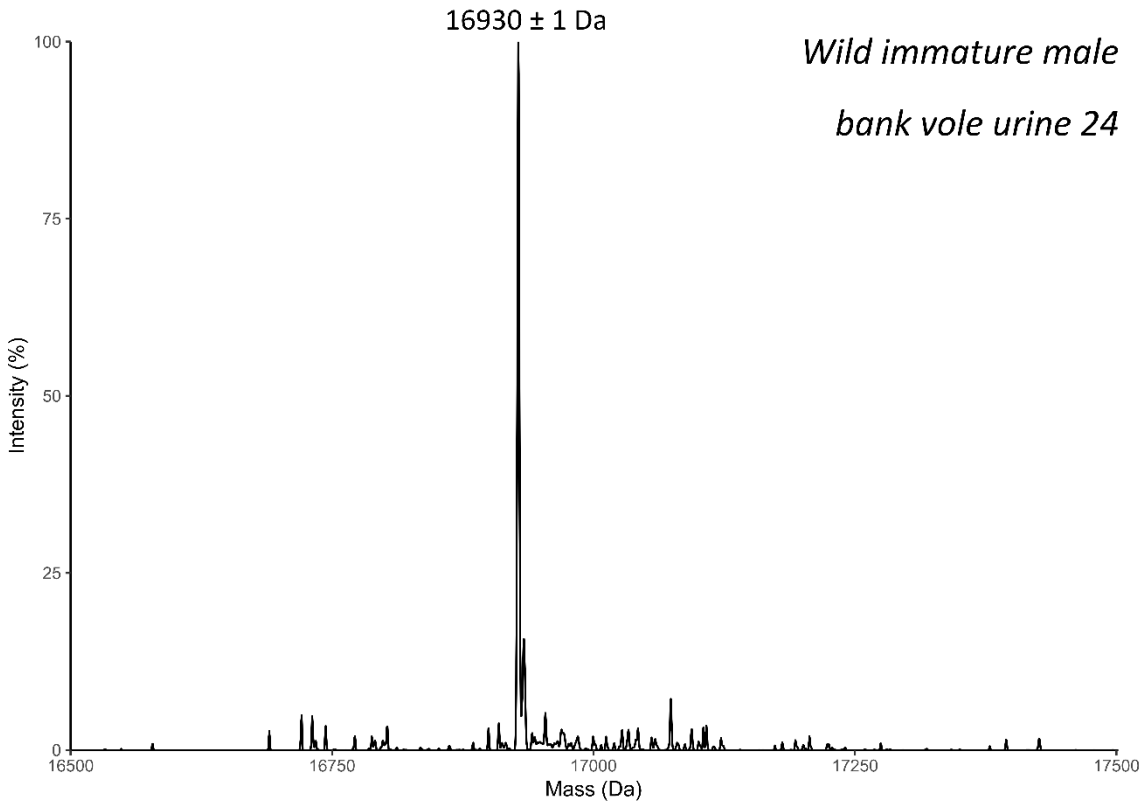


Figure 3.16 Intact mass analysis of urine from wild immature male bank vole 24.

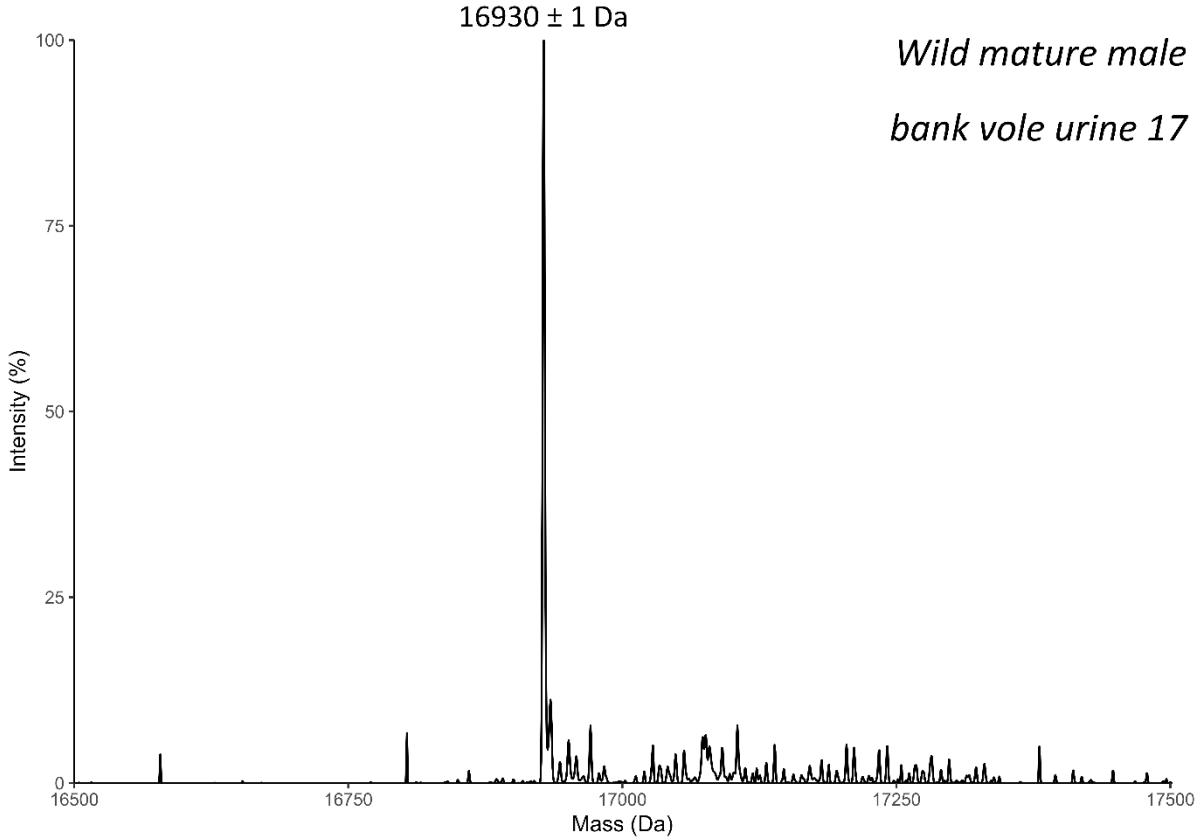


Figure 3.17 Intact mass analysis of urine from wild mature male bank vole 17.

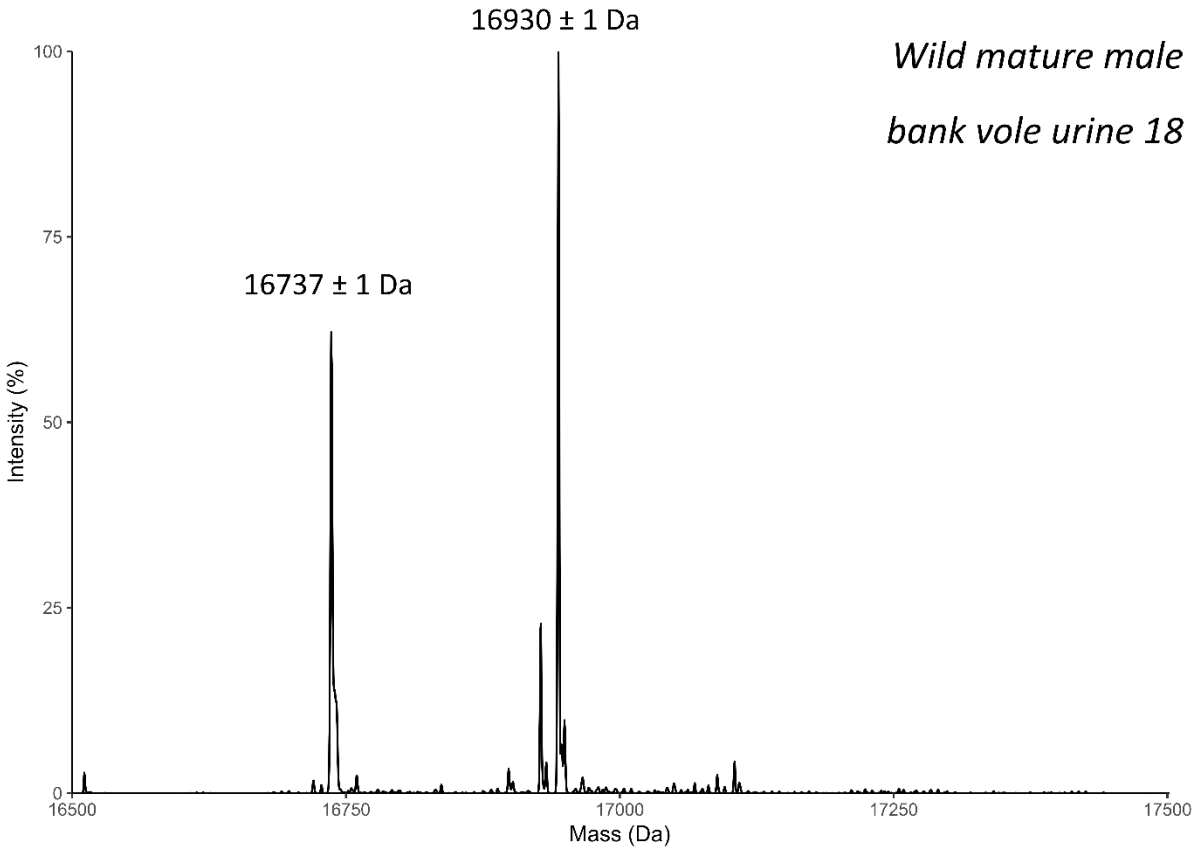


Figure 3.18 Intact mass analysis of urine from wild mature male bank vole 18.

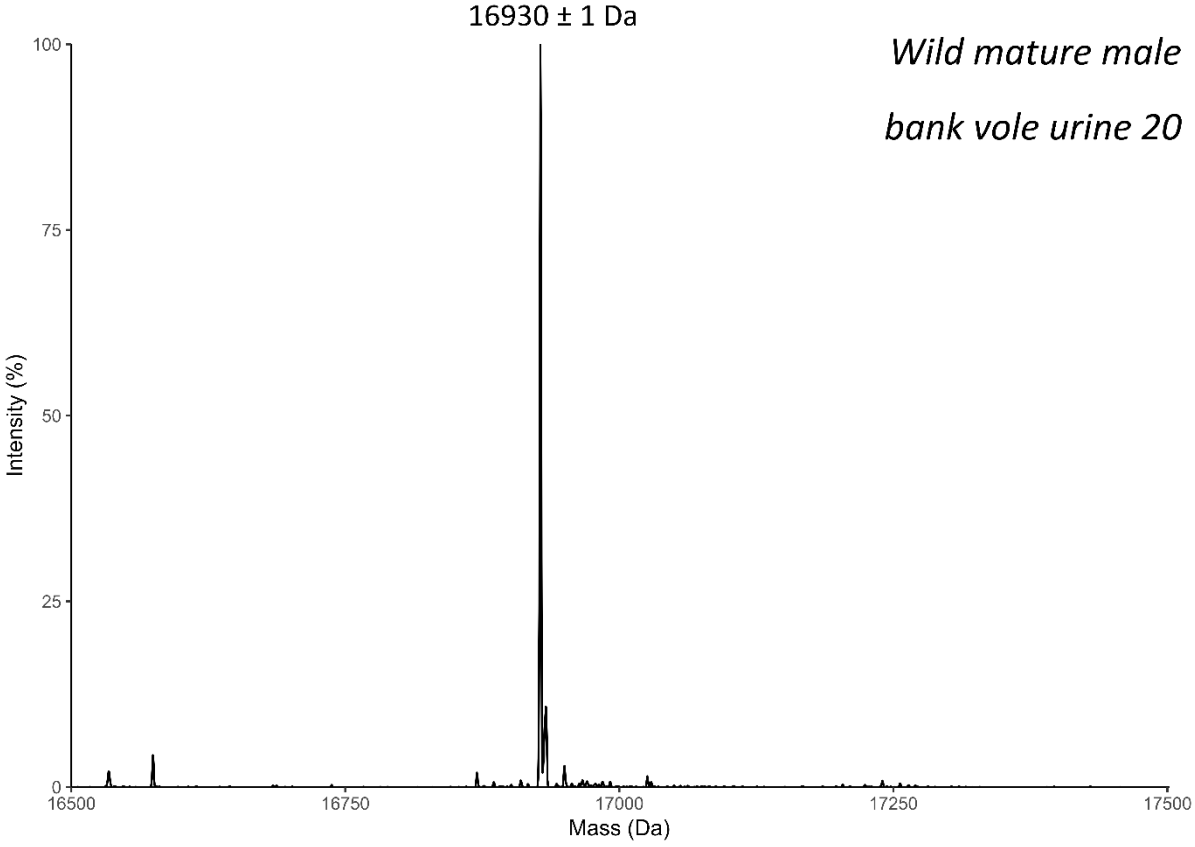


Figure 3.19 Intact mass analysis of urine from wild mature male bank vole 20.

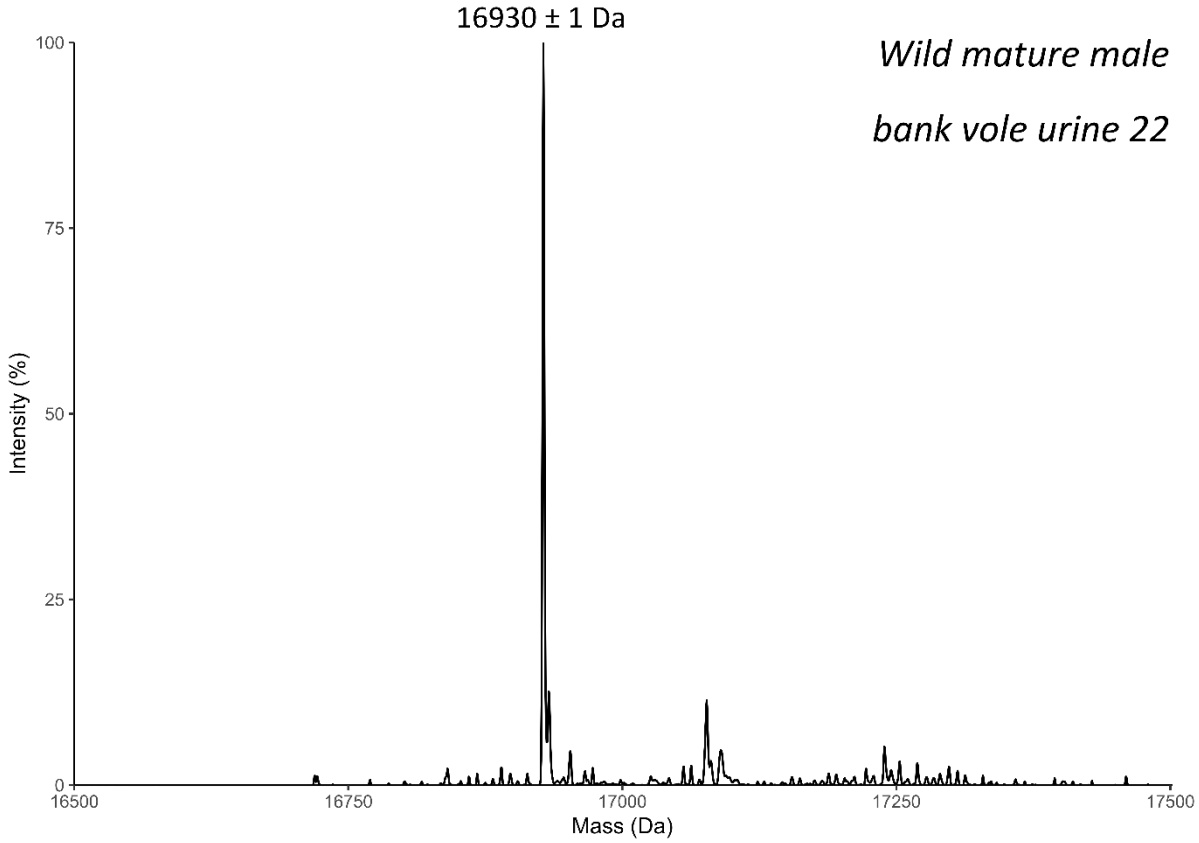


Figure 3.20 Intact mass analysis of urine from wild mature male bank vole 22.

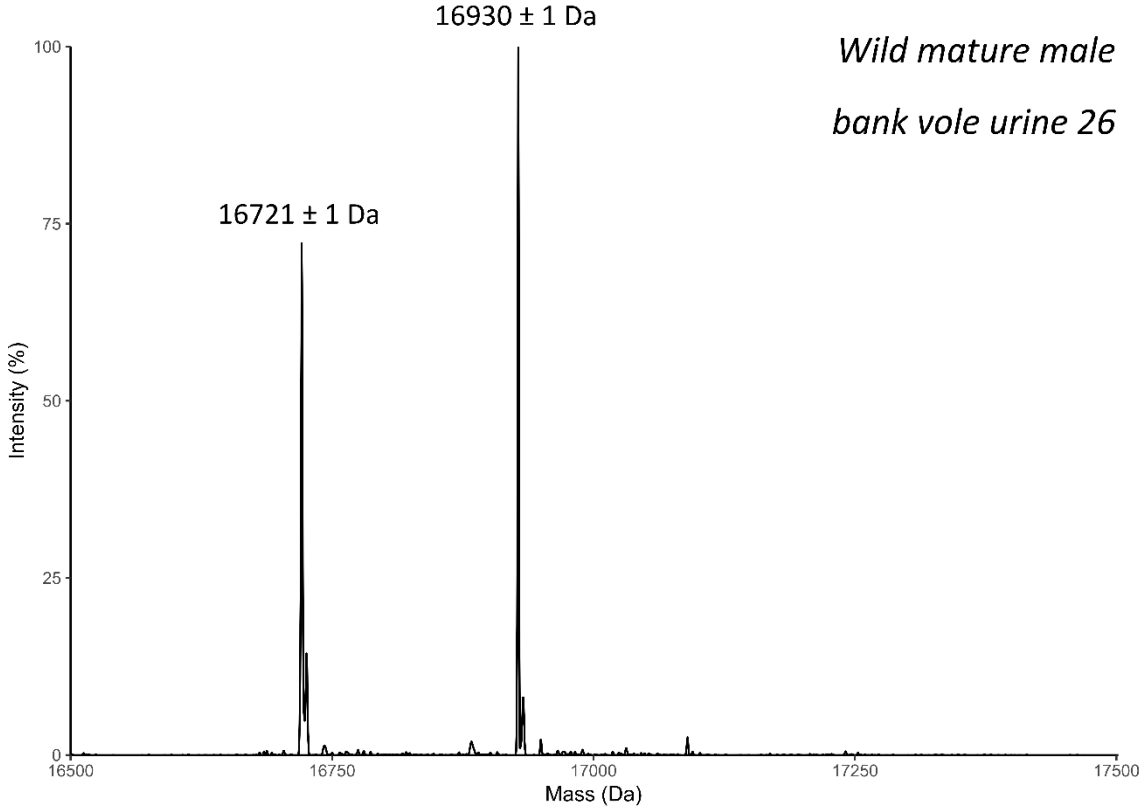


Figure 3.21 Intact mass analysis of urine from wild mature male bank vole 26.

3.6.3 Glareosin sequencing fragment ion data (from paper supplementary)

The following fragment ion data are from the supplementary material of Loxley *et al*. (2017). Each figure legend references the corresponding peptide displayed in Figure 2 of the main paper.

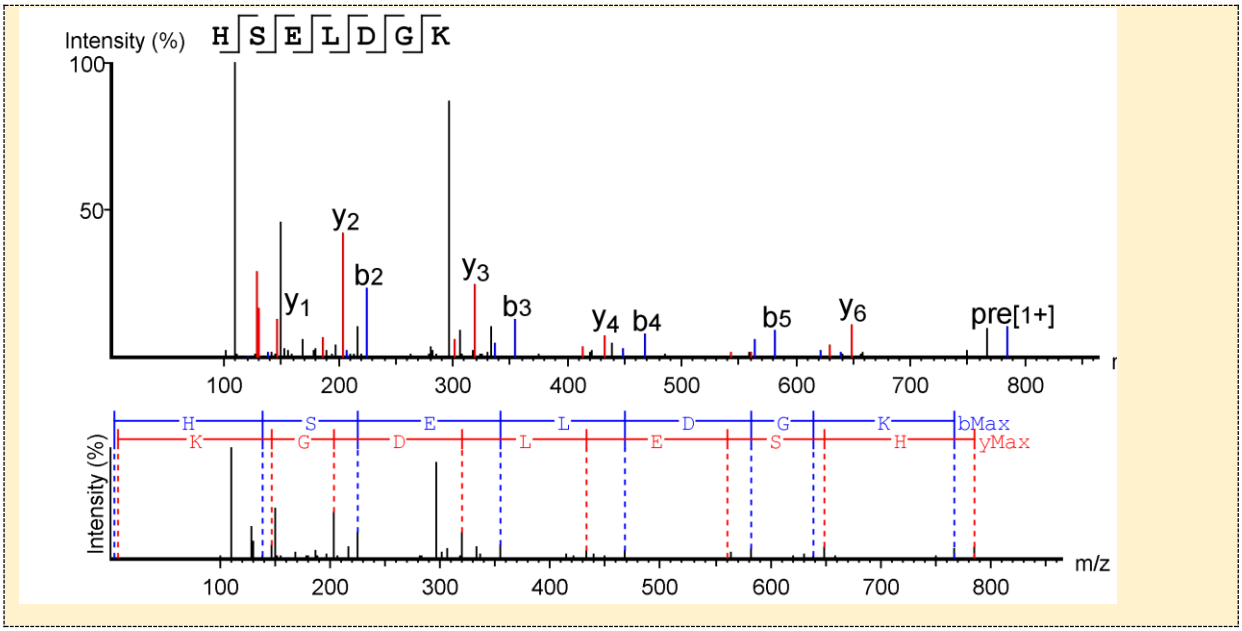


Figure 3.22 *De novo* sequence analysis of the processed MS/MS spectra of *M. glareolus* tryptic peptide 785.3 m/z (Figure 2, t1)

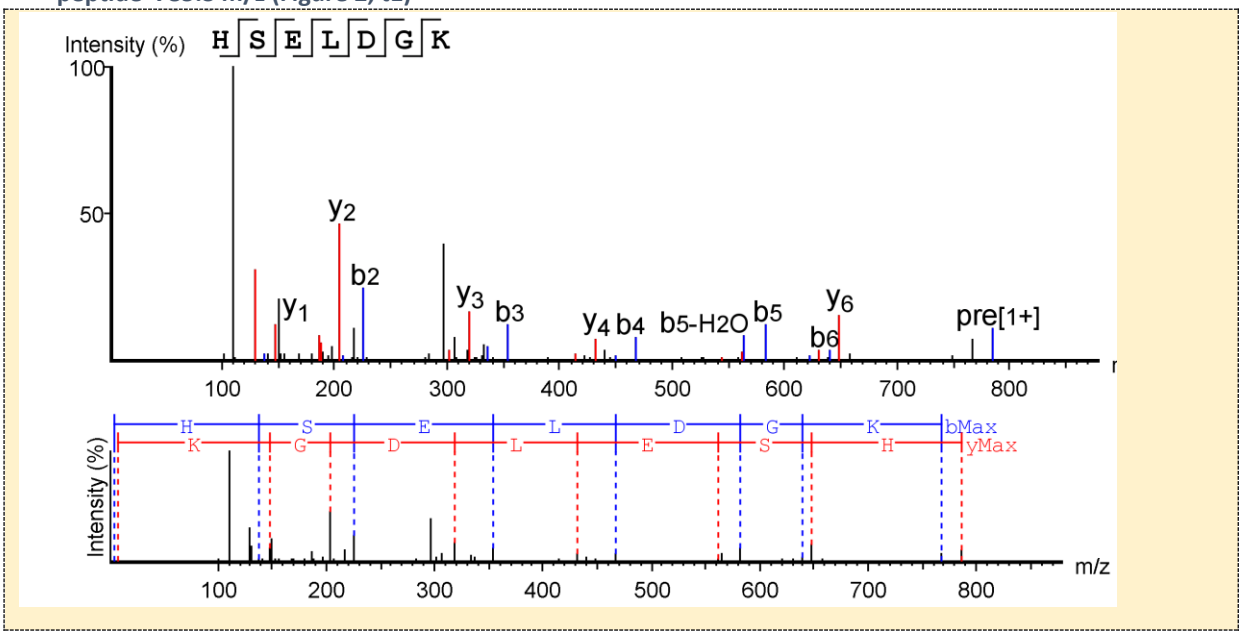


Figure 3.23 *De novo* sequence analysis of the processed MS/MS spectra of *M. glareolus* LysC peptide 785 m/z (Figure 2, t1)

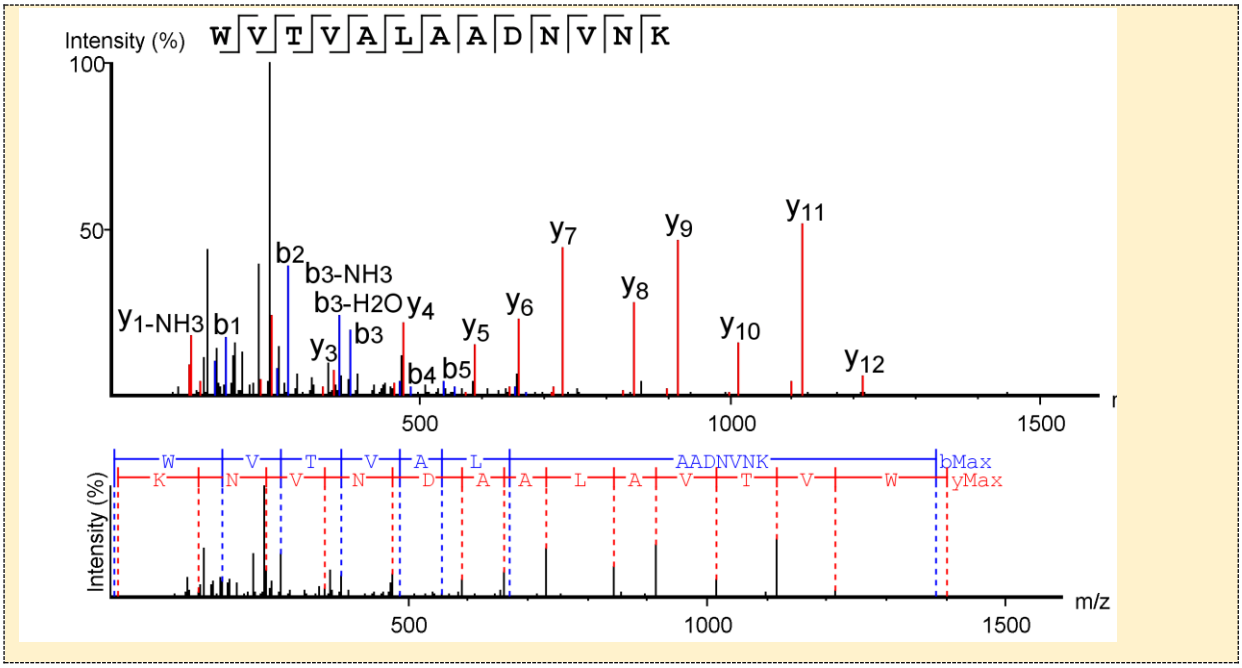


Figure 3.24 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 1400.7 m/z (Figure 2, t2)**

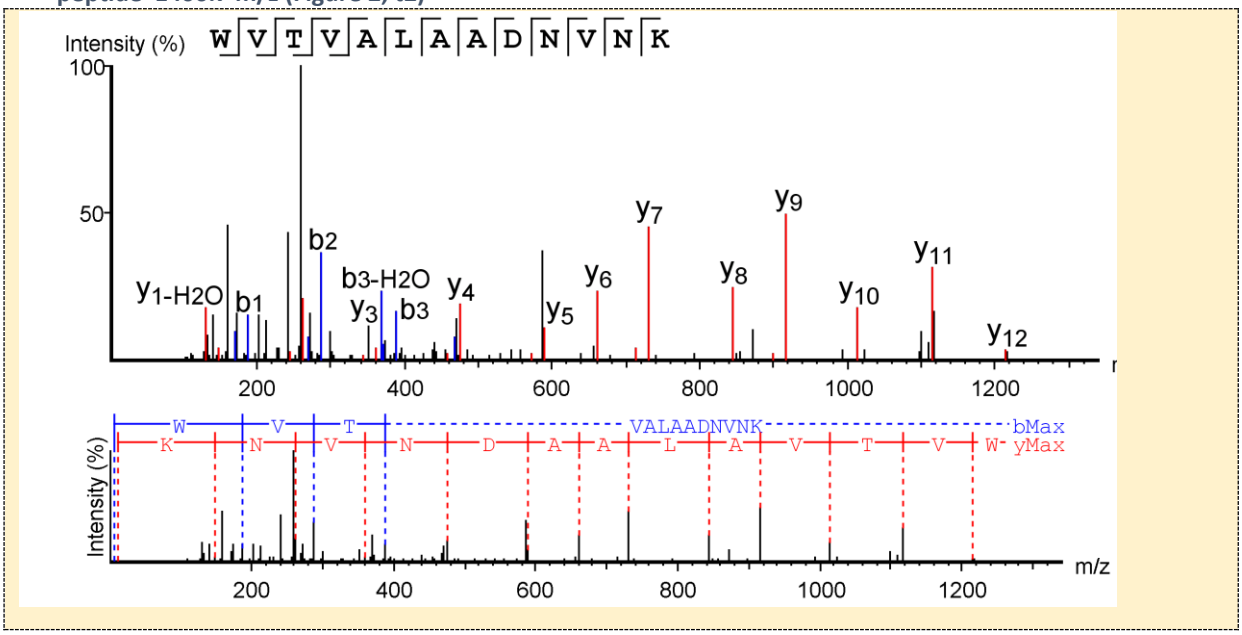


Figure 3.25 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **LysC peptide 1400.7 m/z (Figure 2, t2)**

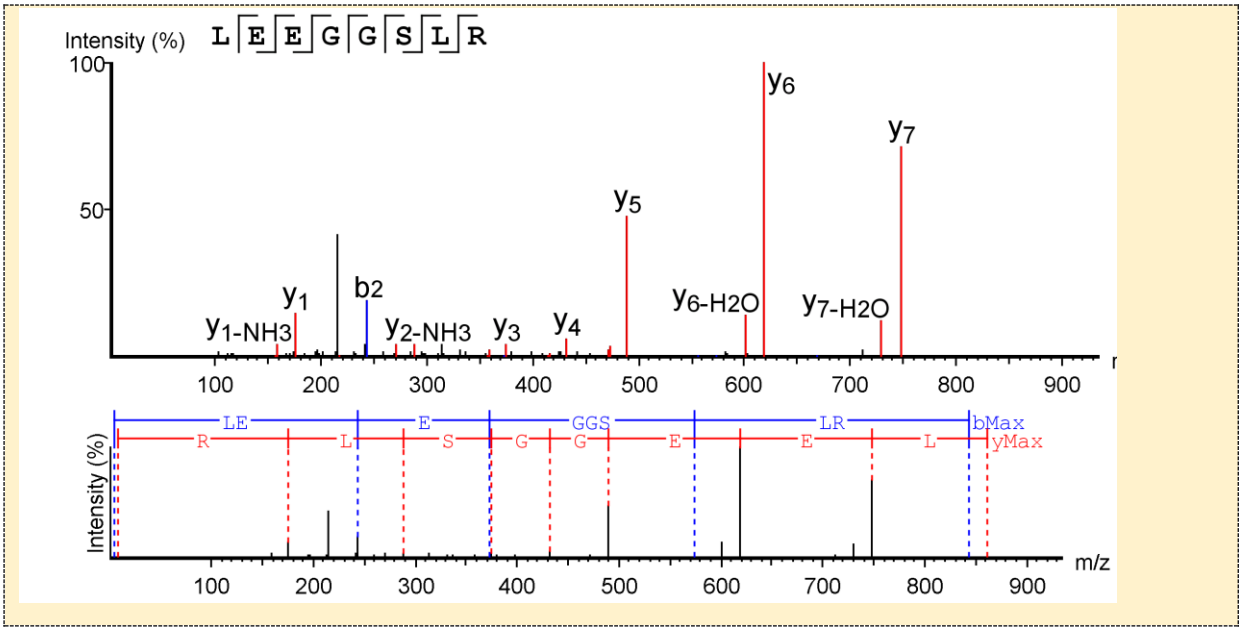


Figure 3.26 *De novo* sequence analysis of the processed MS/MS spectra of *M. glareolus* tryptic peptide 860.4 m/z (Figure 2, t4)

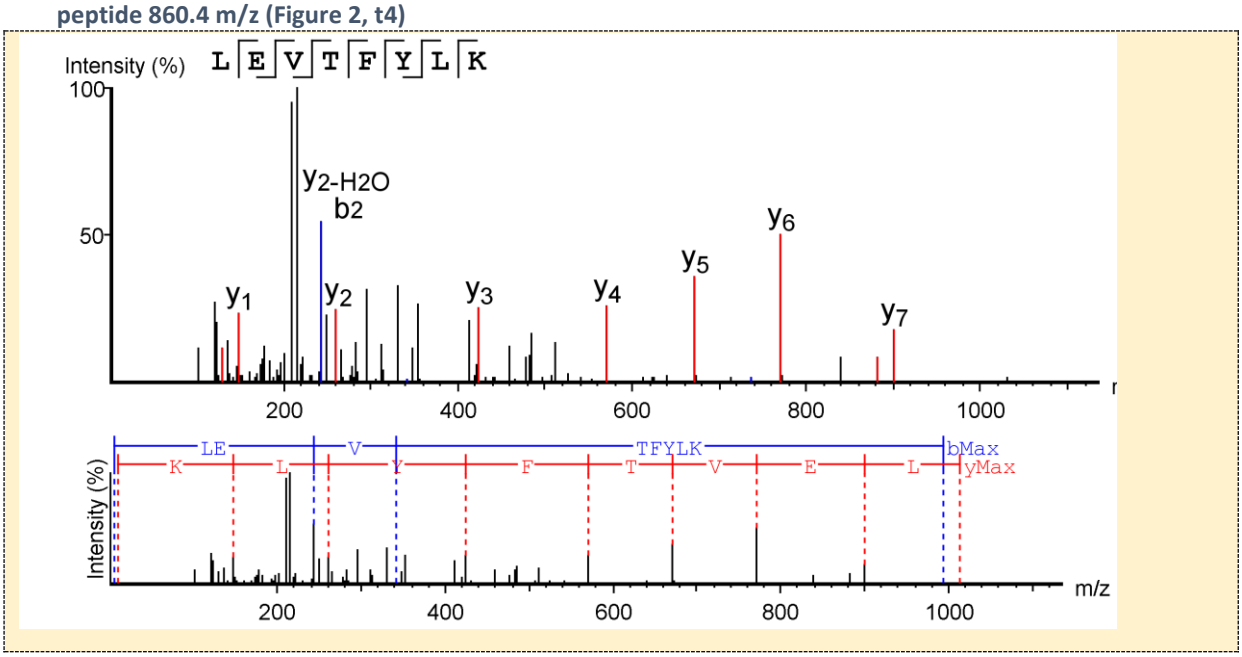


Figure 3.27 *De novo* sequence analysis of the processed MS/MS spectra of *M. glareolus* tryptic peptide 1012.5 m/z (Figure 2, t8)

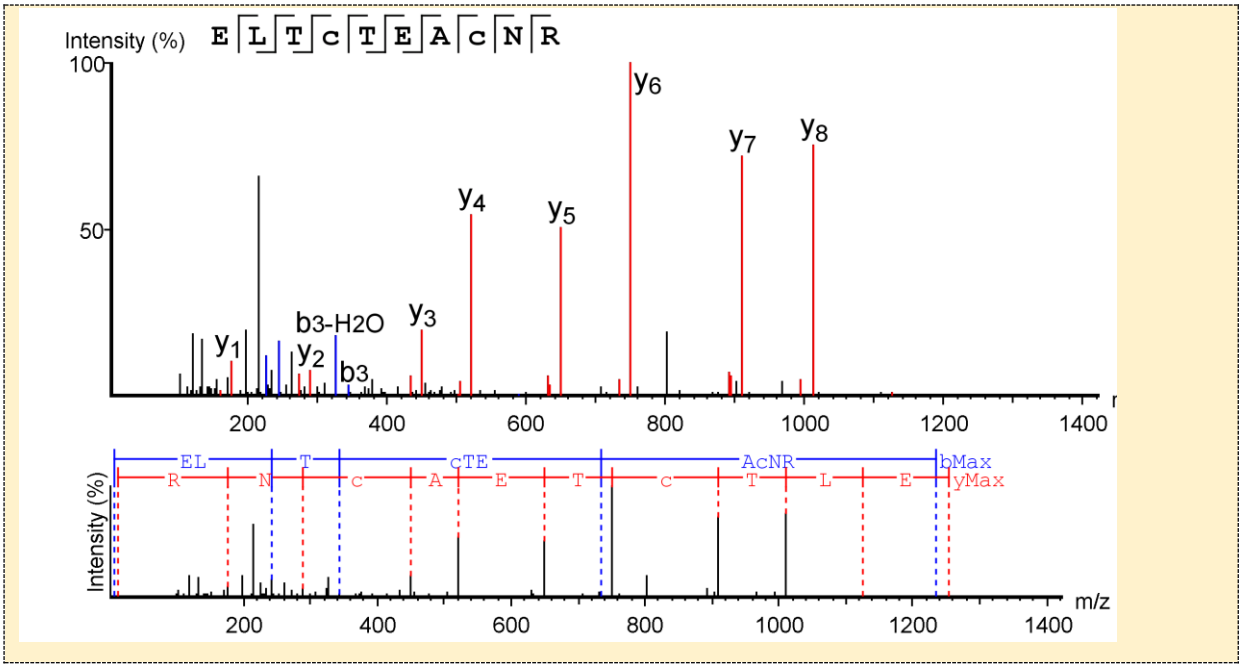


Figure 3.28 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 1253.5 m/z (Figure 2, t5)**

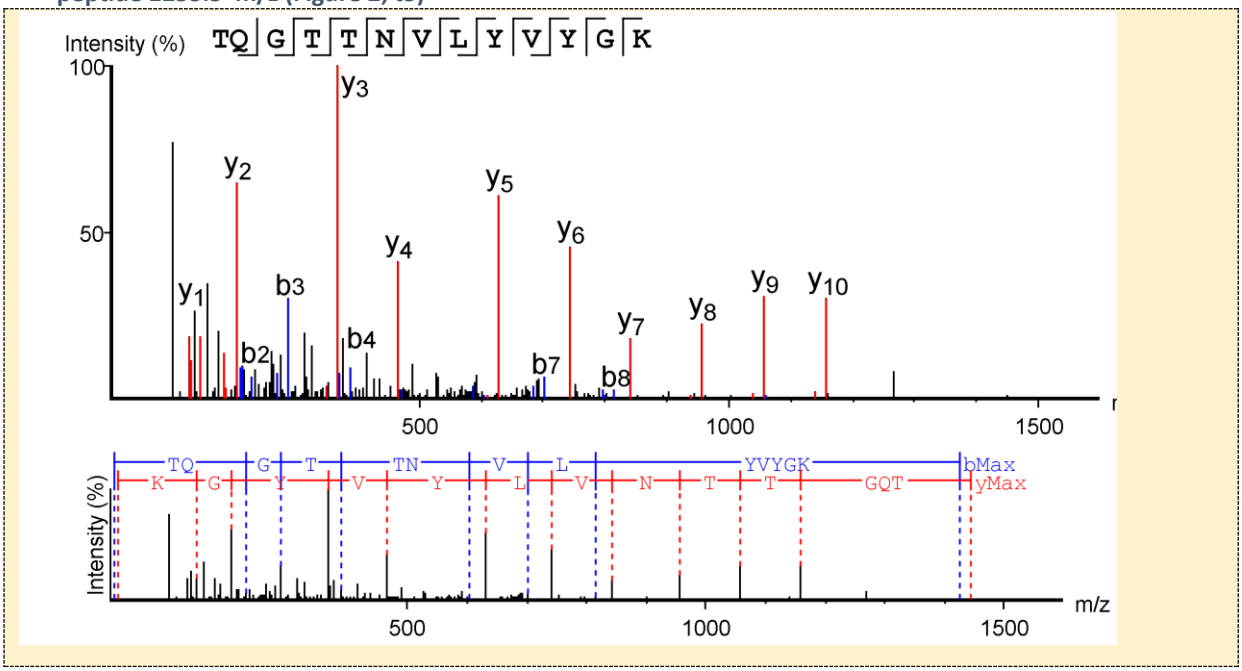


Figure 3.29 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 1443.7 m/z (Figure 2, t11)**

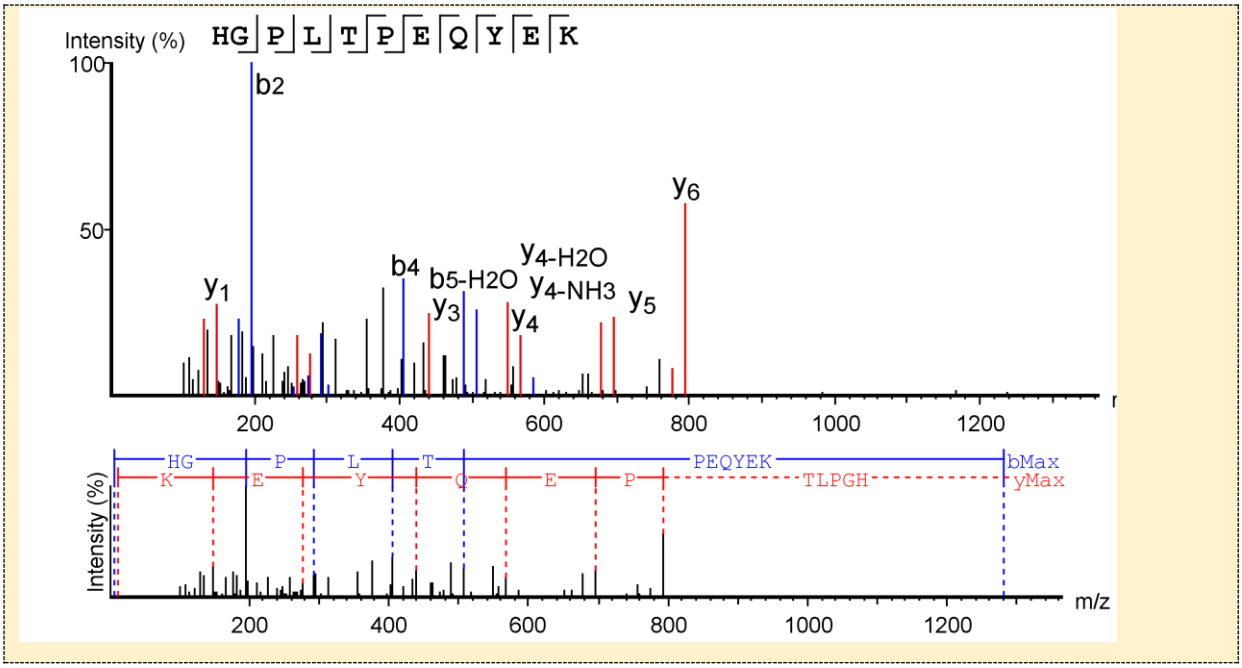


Figure 3.30 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 1298.6 m/z(Figure 2, t12)**

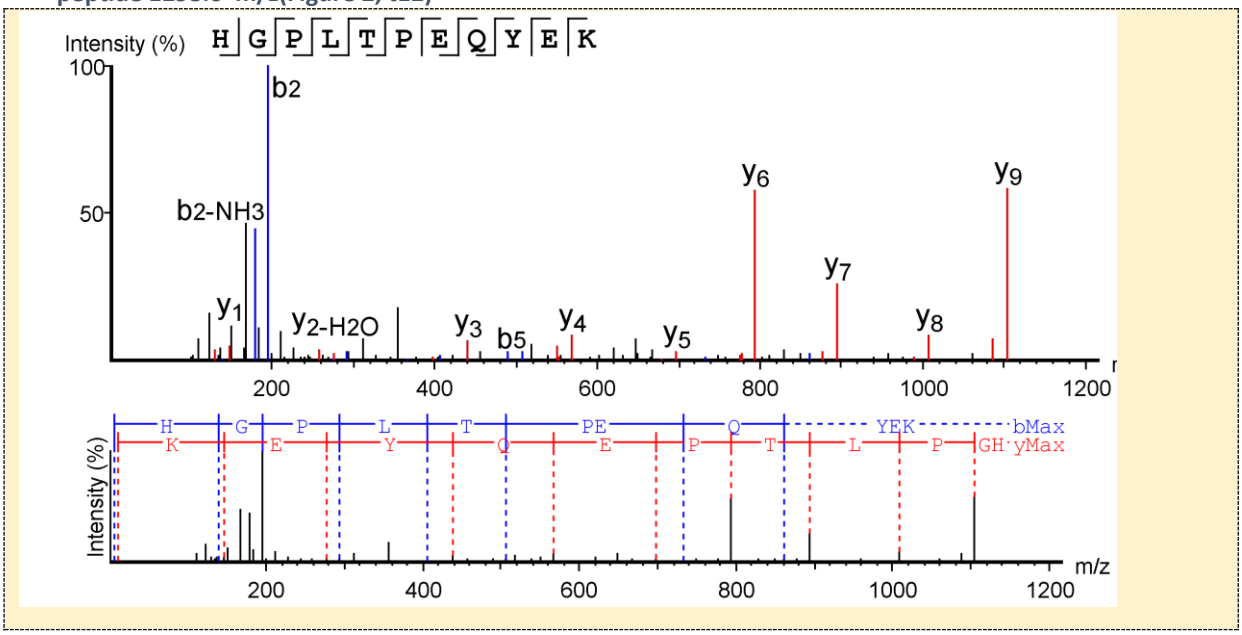


Figure 3.31 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **LysC peptide 1298.6 m/z (Figure 2, t12)**

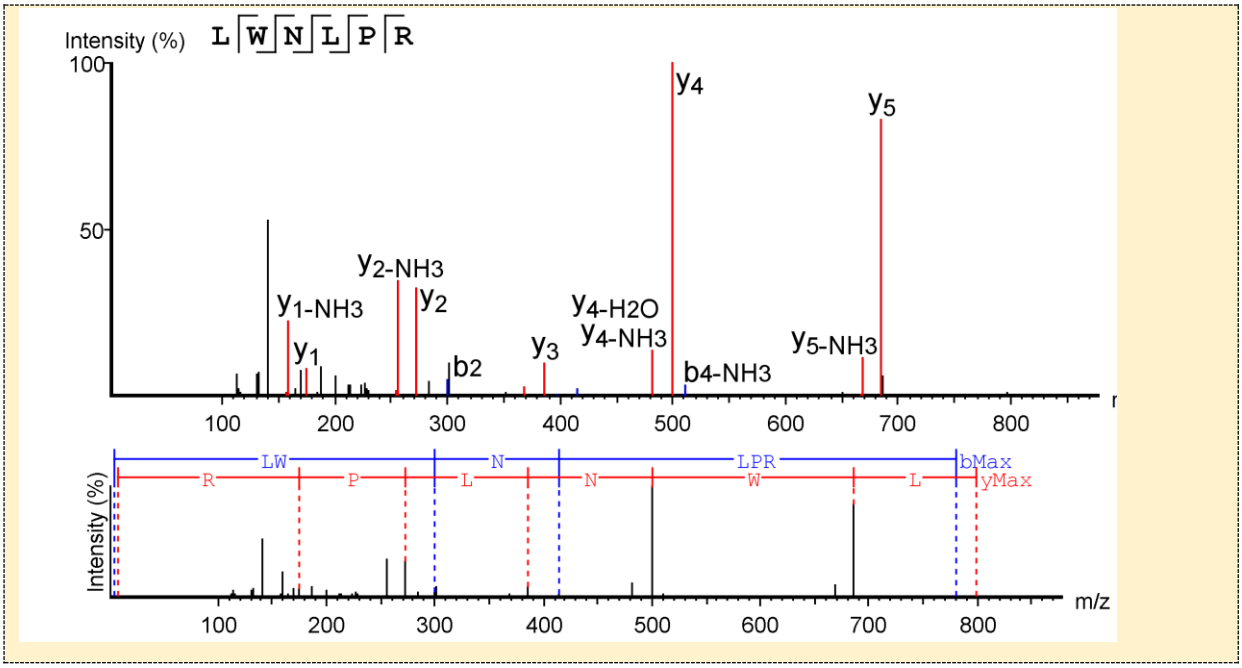


Figure 3.32 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 798.4 m/z (Figure 2, t15)**

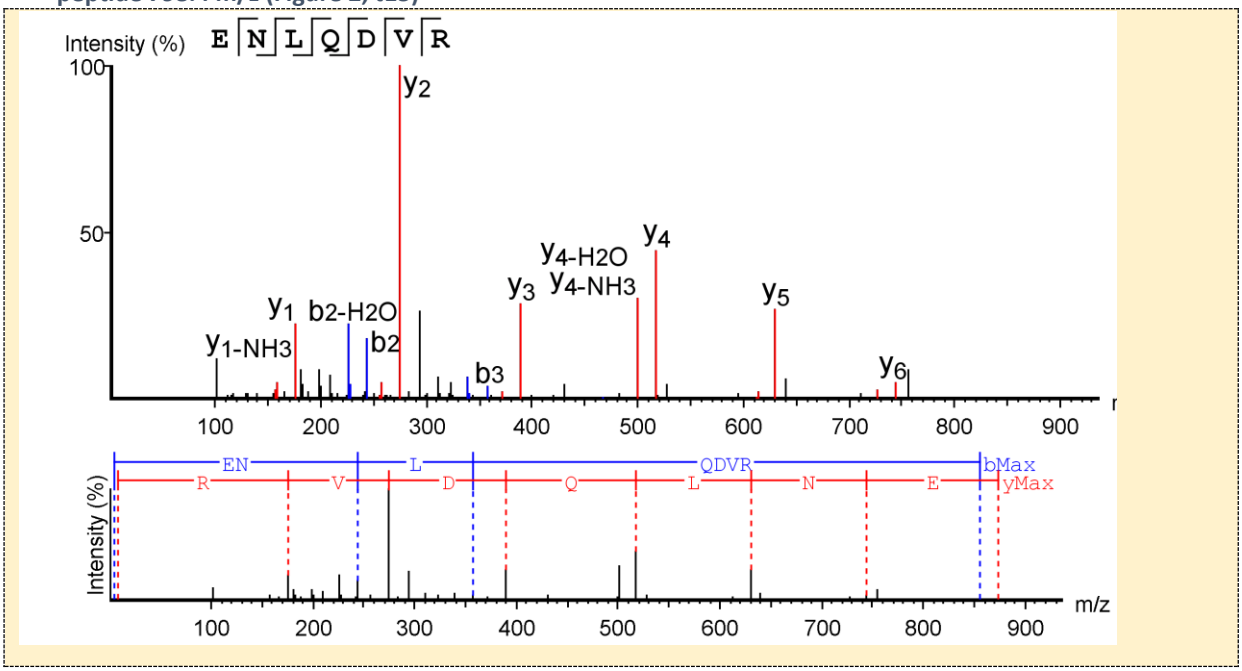


Figure 3.33 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 873.4 m/z (Figure 2, t15)**

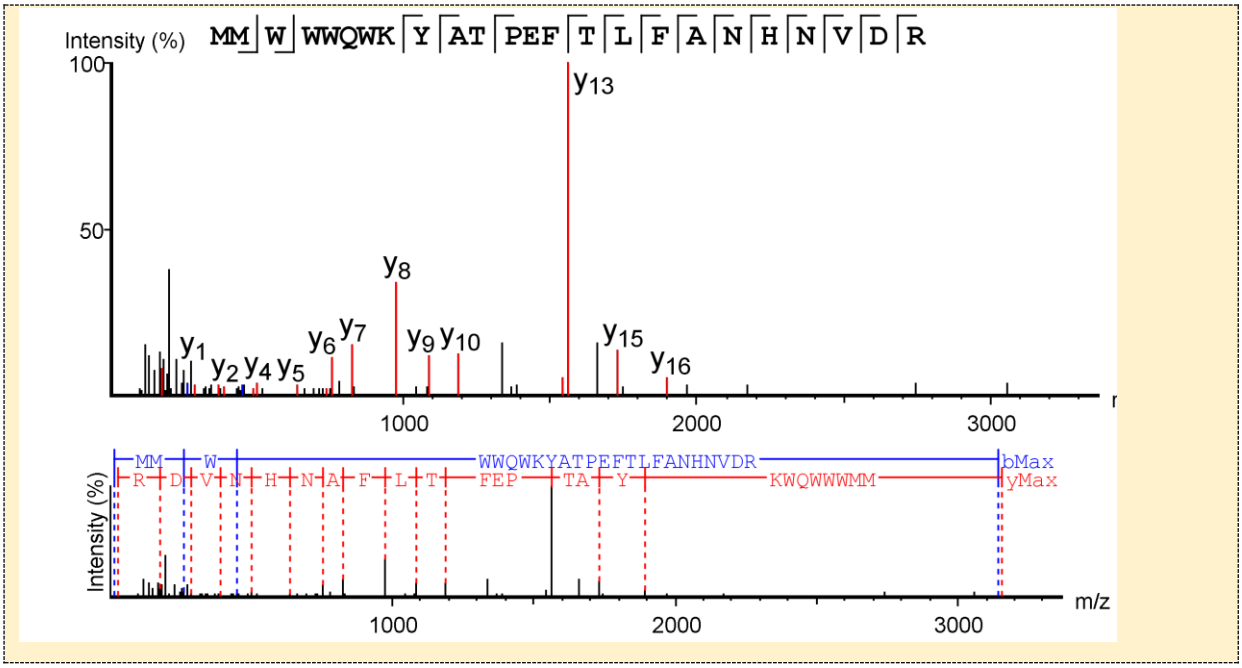


Figure 3.34 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 3157.4 m/z (Figure 2, t9)**

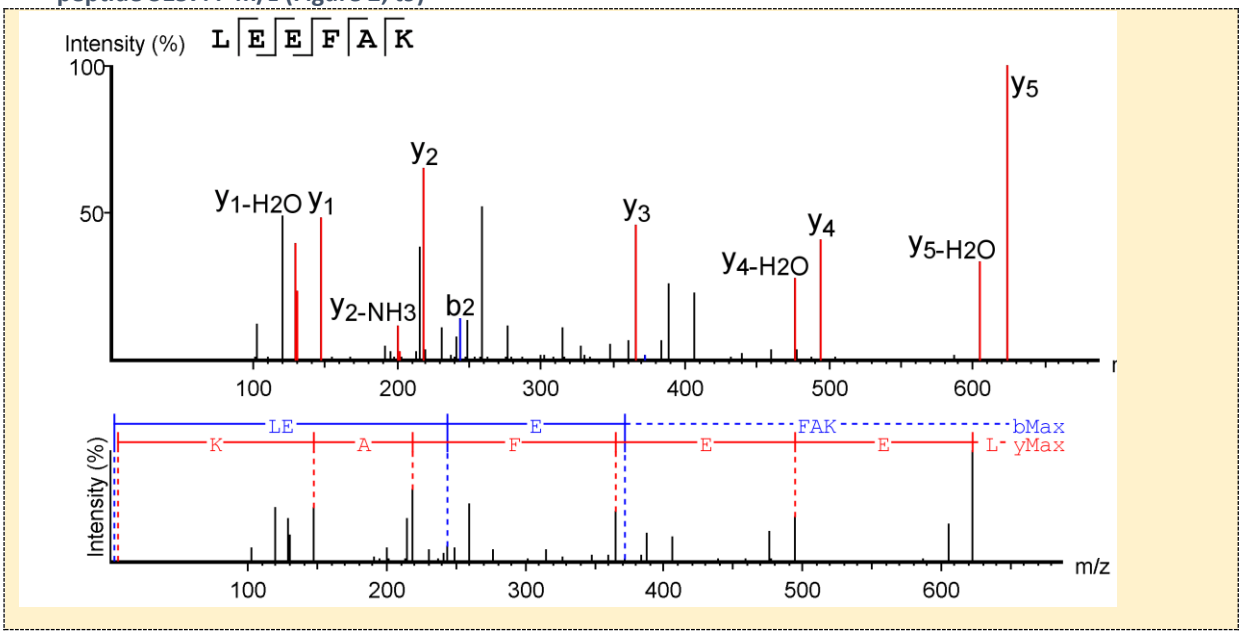


Figure 3.35 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **tryptic peptide 736.3 m/z (Figure 2, t13)**

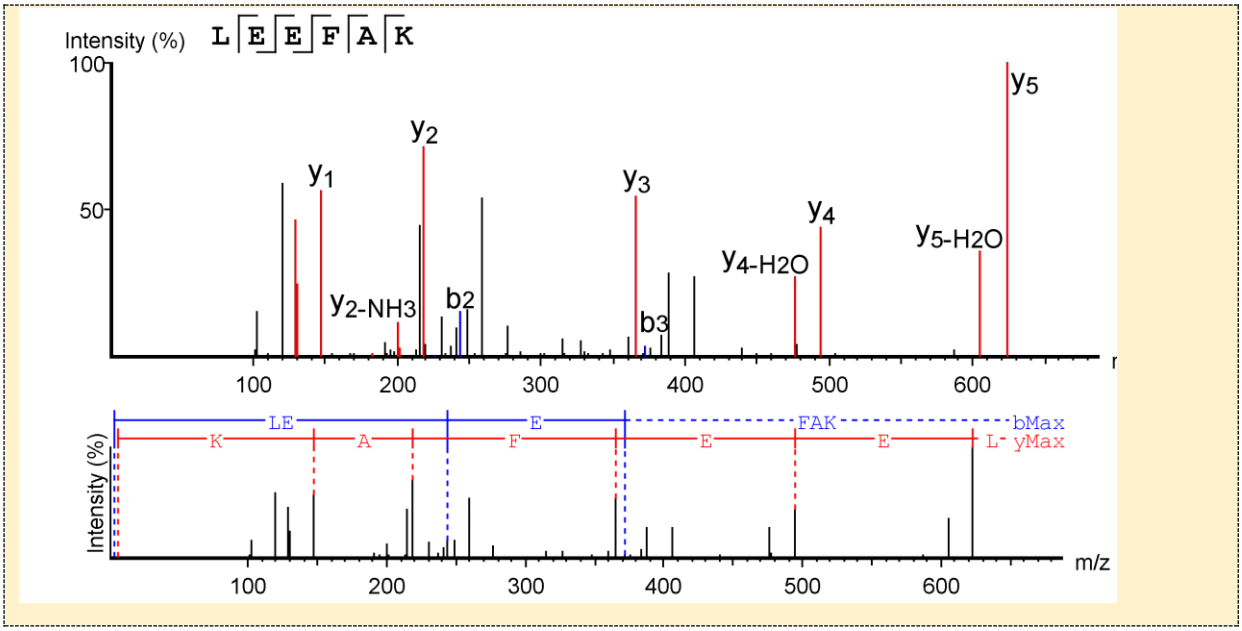


Figure 3.36 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **LysC peptide 736.3 m/z (Figure 2, t13)**

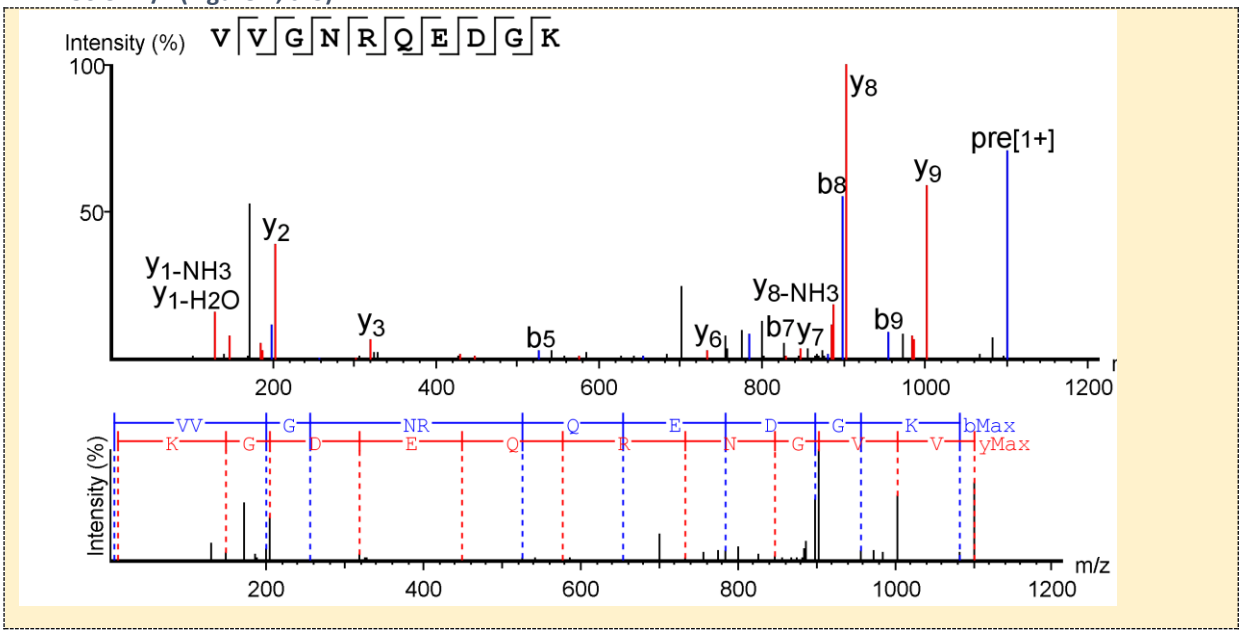


Figure 3.37 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **LysC peptide 1101.5m/z (Figure 2, c1)**

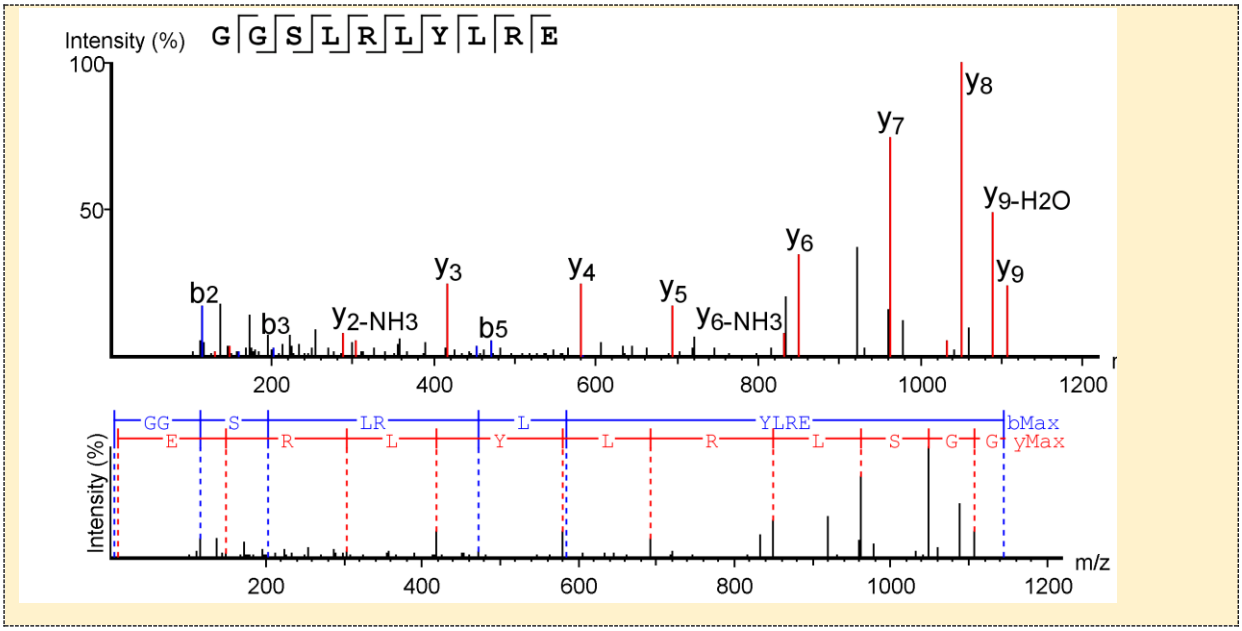


Figure 3.38 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **GluC peptide 1163.6 m/z (Figure 2, g2)**

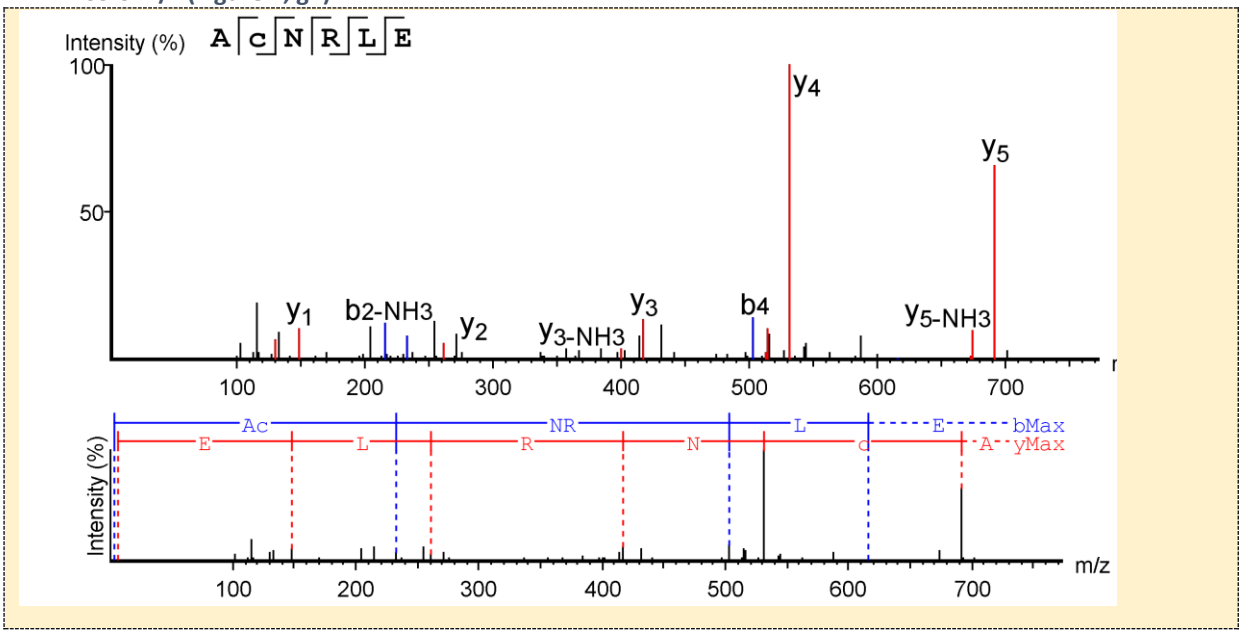


Figure 3.39 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **GluC peptide 762.3 m/z (Figure 2, g1)**

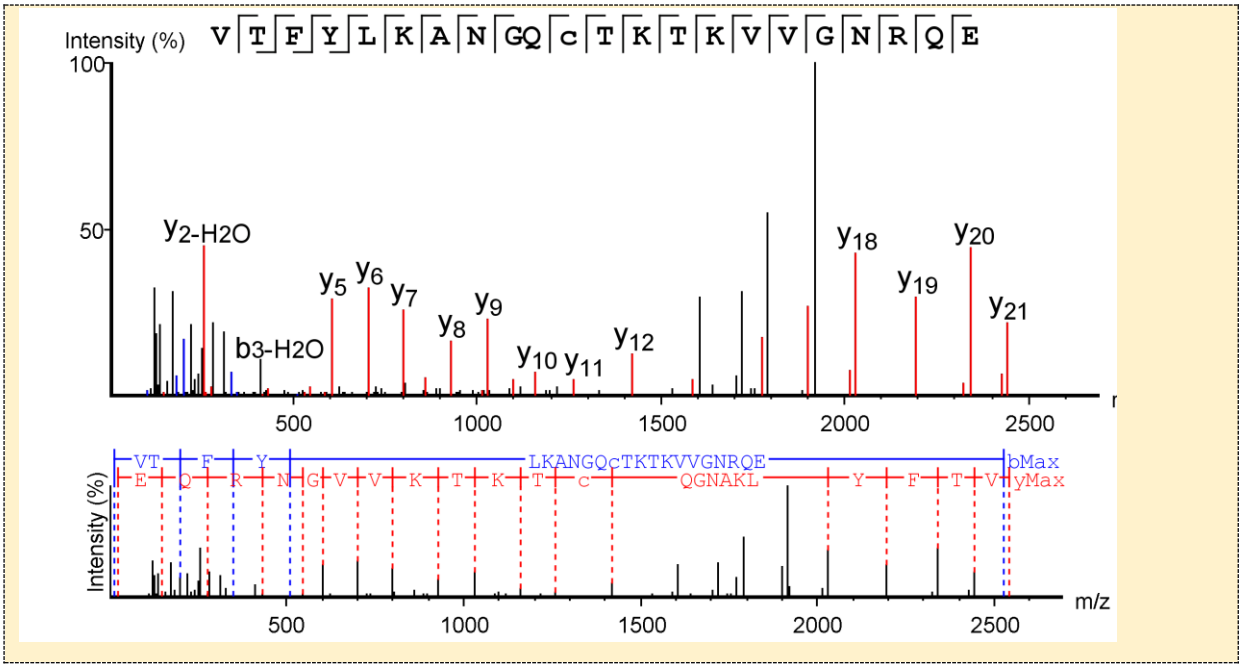


Figure 3.40 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **GluC peptide 2541.3 m/z (Figure 2, g1)**

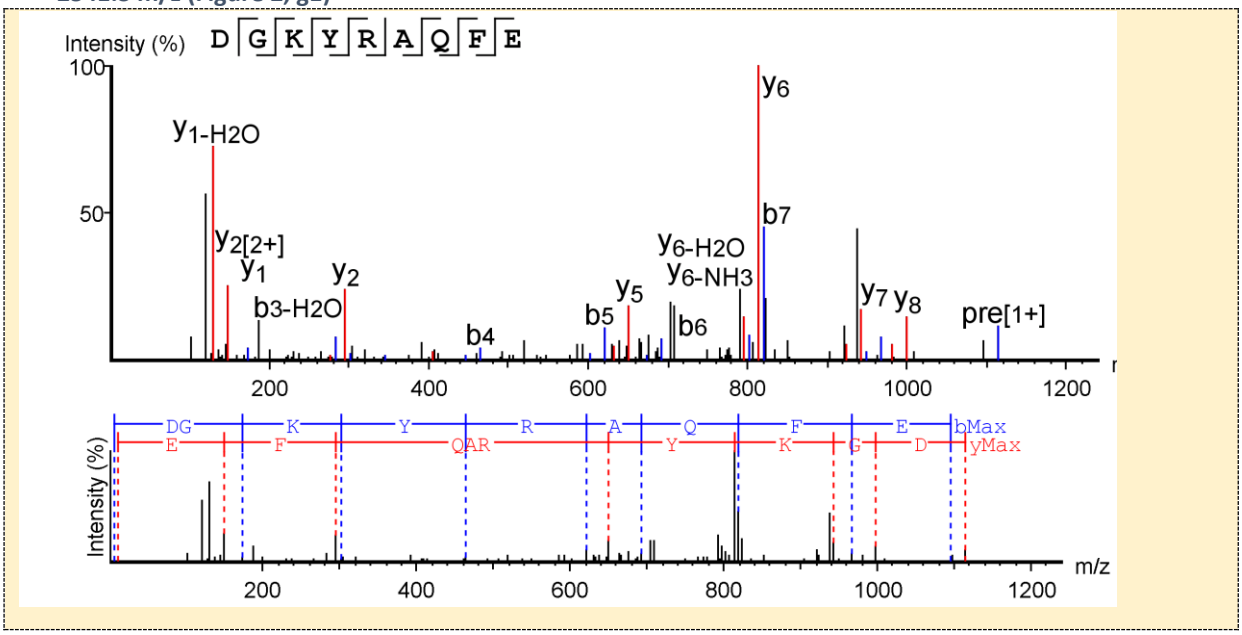


Figure 3.41 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **GluC peptide 1113.5 m/z (Figure 2, g5)**

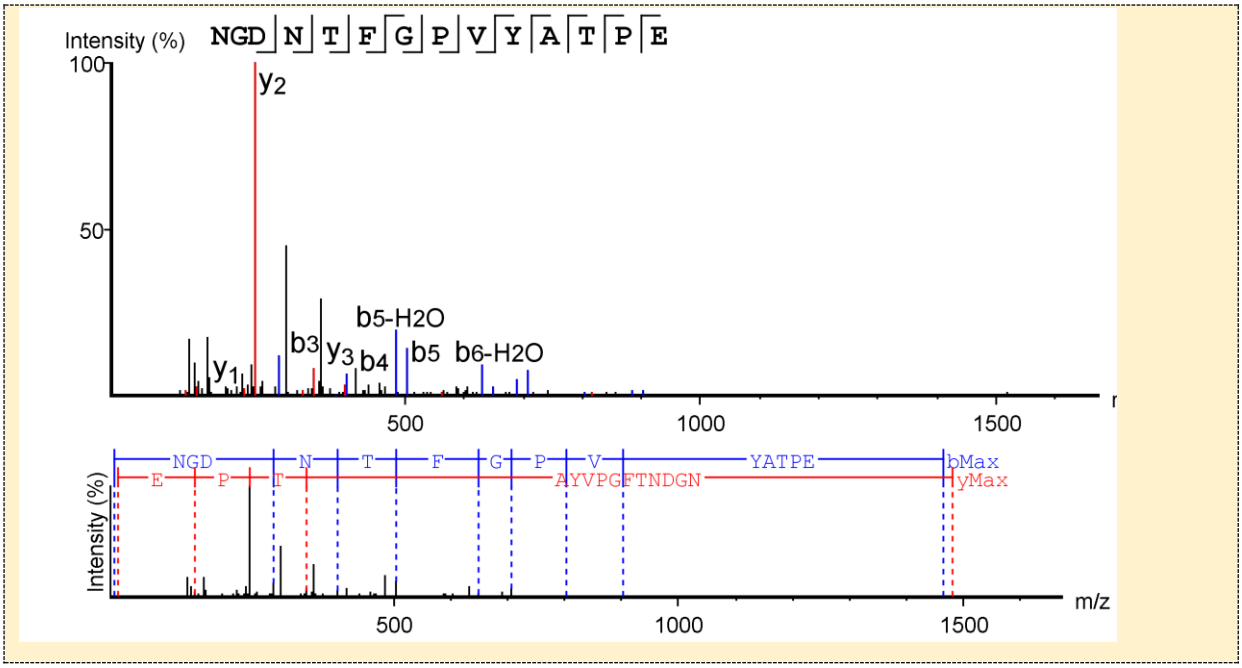


Figure 3.42 *De novo* sequence analysis of the processed MS/MS spectra of *M. glareolus* GluC peptide 1481.6 m/z (Figure 2, g61)

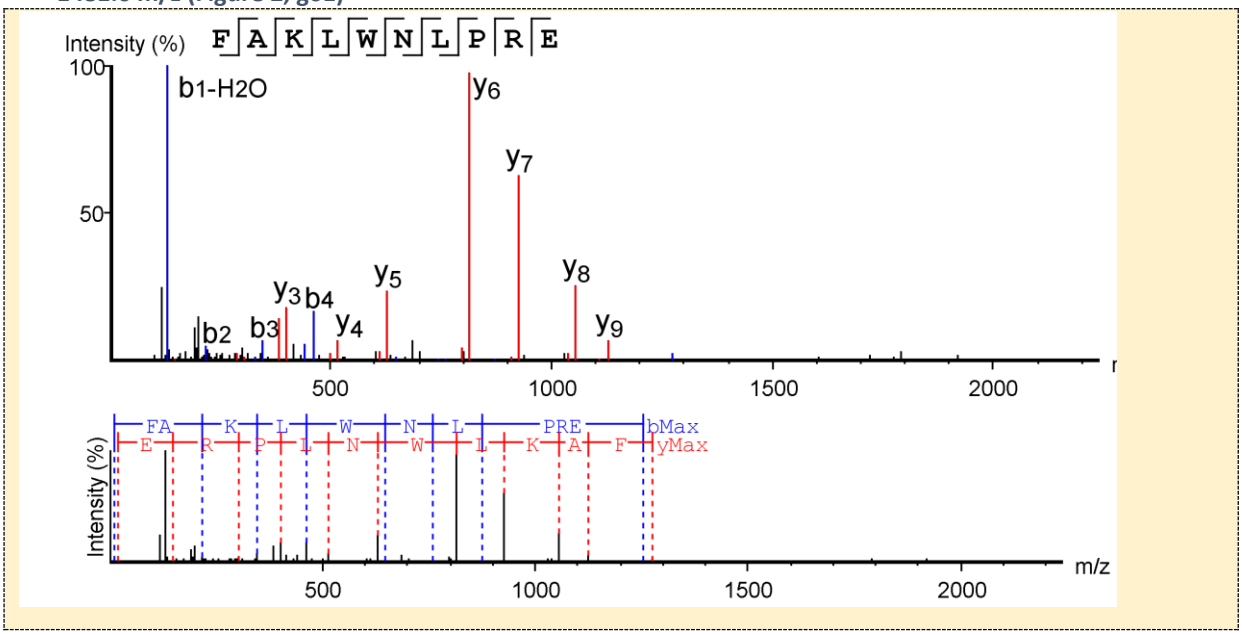


Figure 3.43 *De novo* sequence analysis of the processed MS/MS spectra of *M. glareolus* GluC peptide 1273.7 m/z (Figure 2, g8)

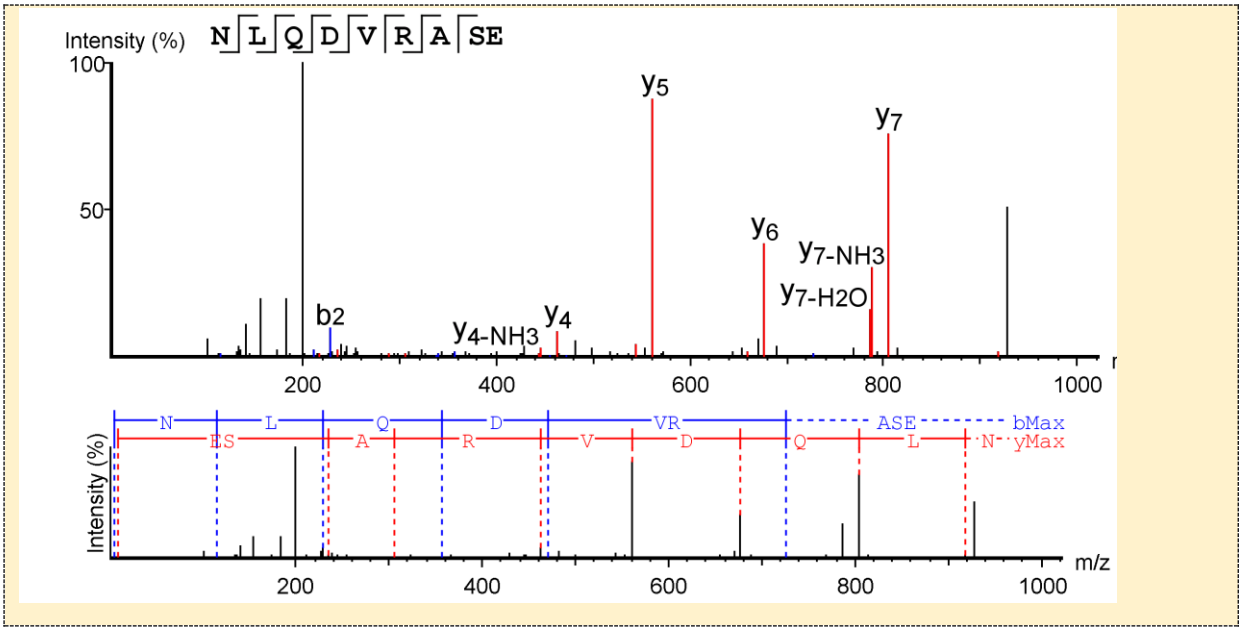


Figure 3.44 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. glareolus*** **GluC peptide 1031.5 m/z (Figure 2, g9)**

3.6.4 *Glareosin heavy leucine data*

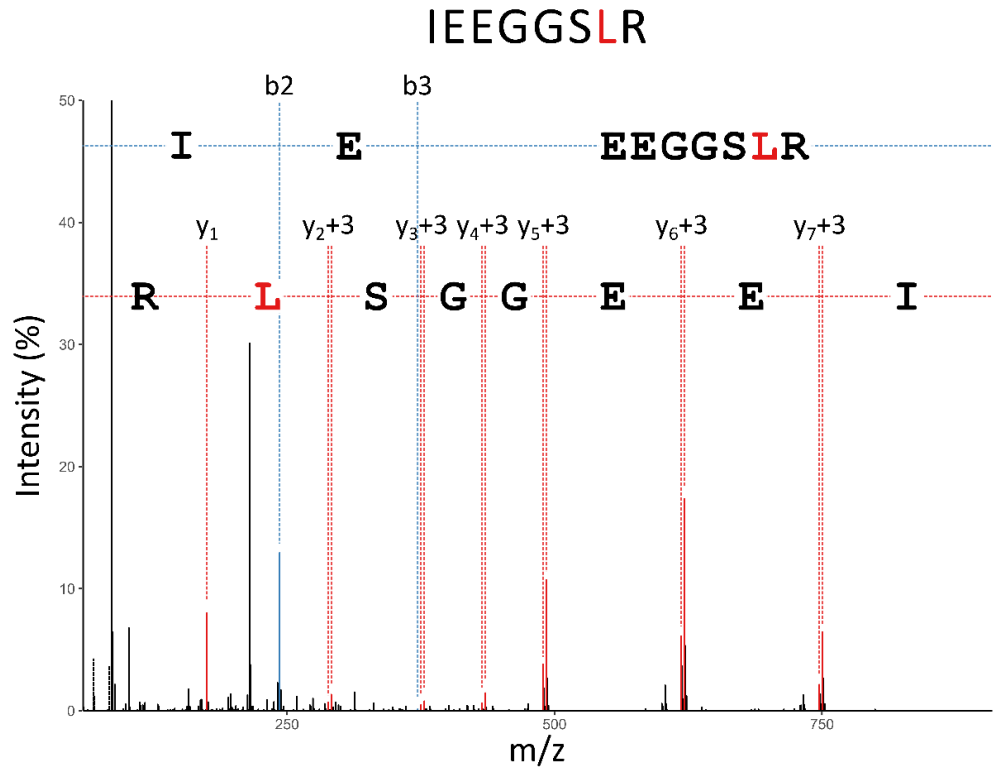


Figure 3.45 Leucine & isoleucine determination from the heavy labelled, processed MS/MS spectra of *M. glareolus* trypsin peptide IEEGGSLR 430.7 m/z (Figure 3.11).

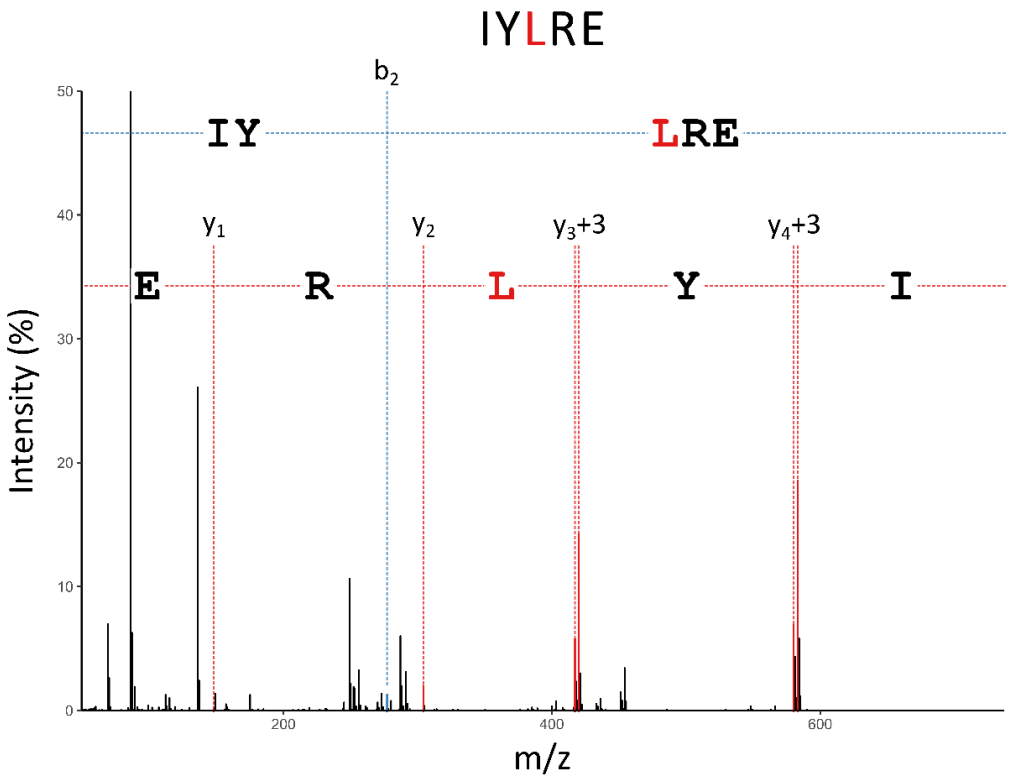


Figure 3.46 Leucine & isoleucine determination from the heavy labelled, processed MS/MS spectra of *M. glareolus* trypsin peptide IYLRE 693.4 m/z (Figure 3.11).

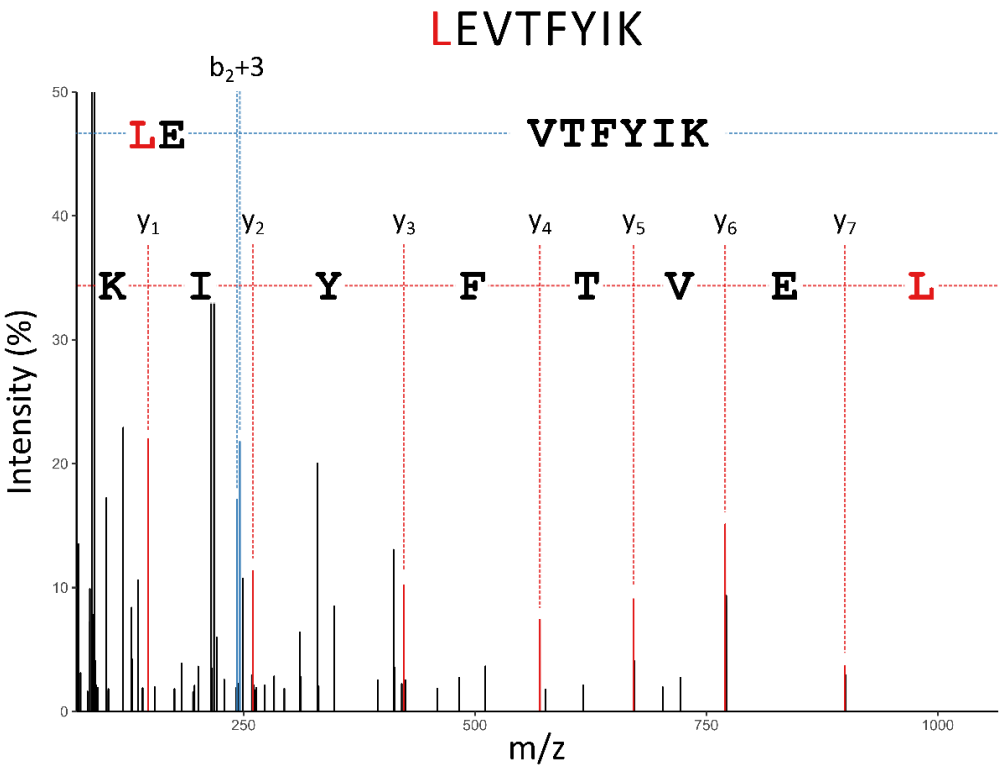


Figure 3.47 Leucine & isoleucine determination from the heavy labelled, processed MS/MS spectra of *M. glareolus* trypsin peptide LEVTFYIK 385.7 m/z (Figure 3.11).

3.6.5 *Multiple sequence alignment for phylogenetic analysis (from paper supplementary)*

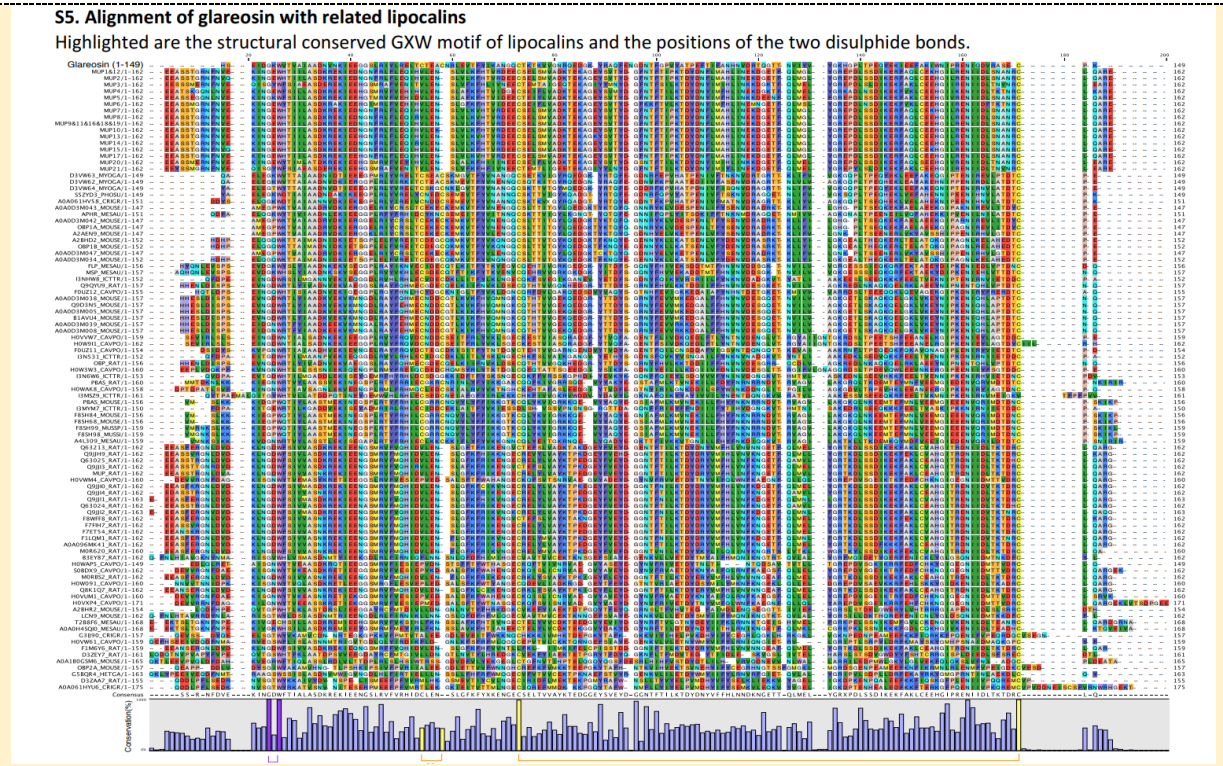


Figure 3.48 Multiple sequence alignment of glareosin with related lipocalins.

3.6.6 *Bank Vole Bladder Urine Intact Protein Mass Spectra*

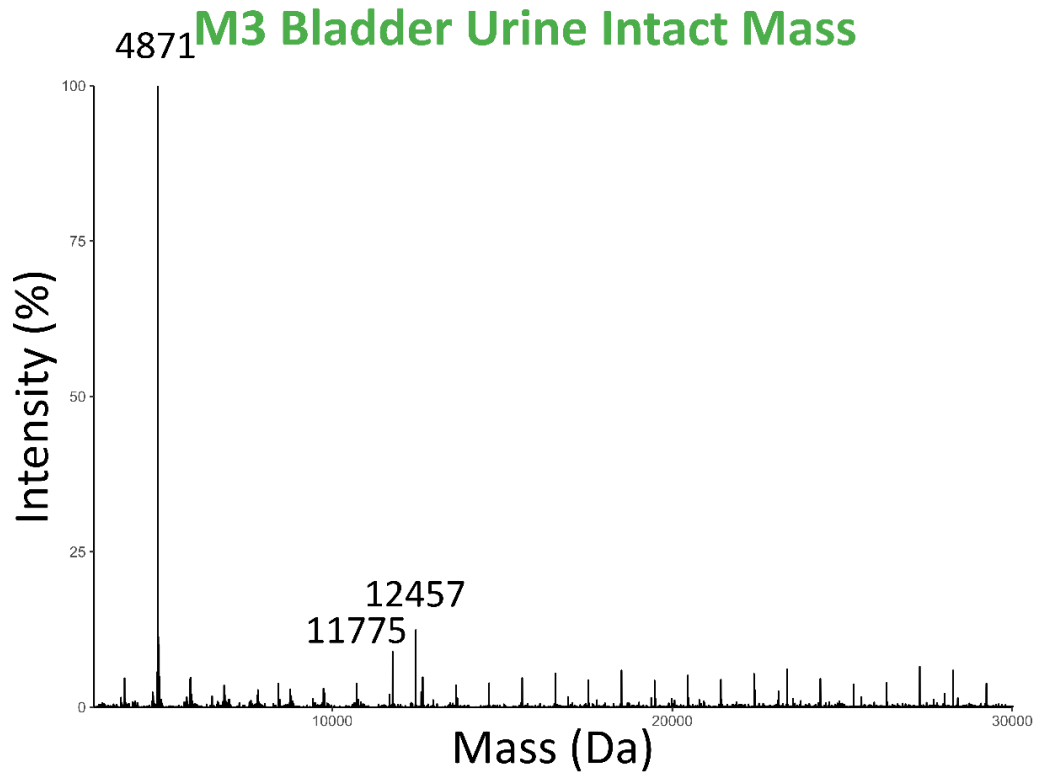


Figure 3.49 Intact mass of bank vole bladder urine from male 3.

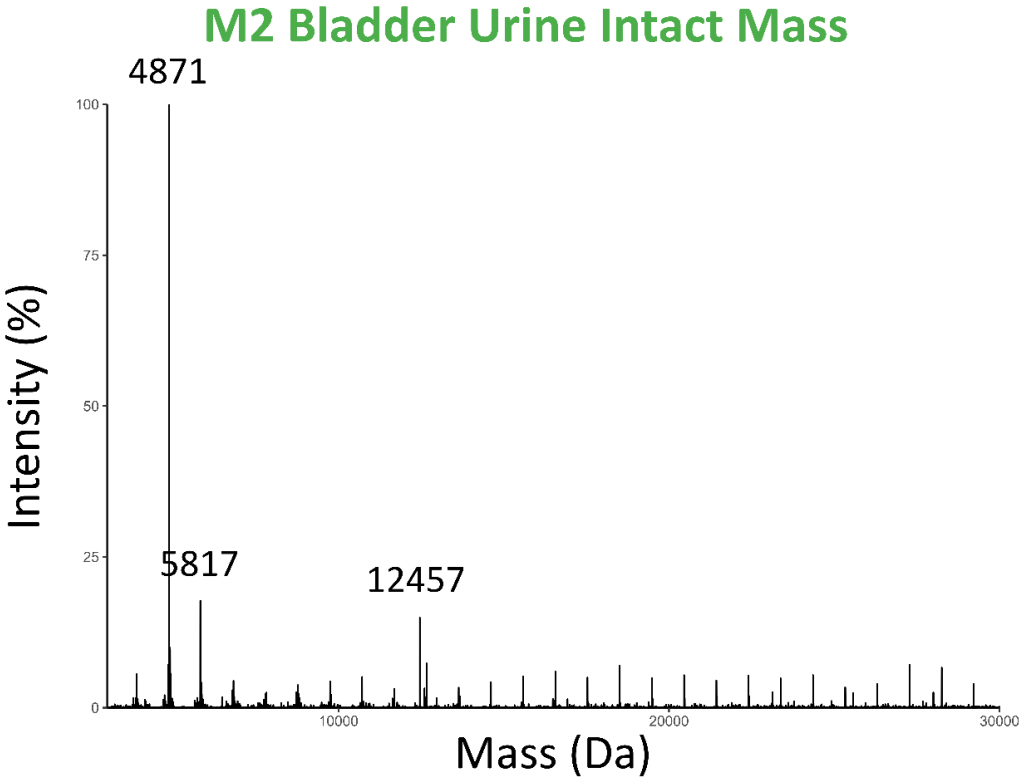


Figure 3.50 Intact mass of bank vole bladder urine from male 2.

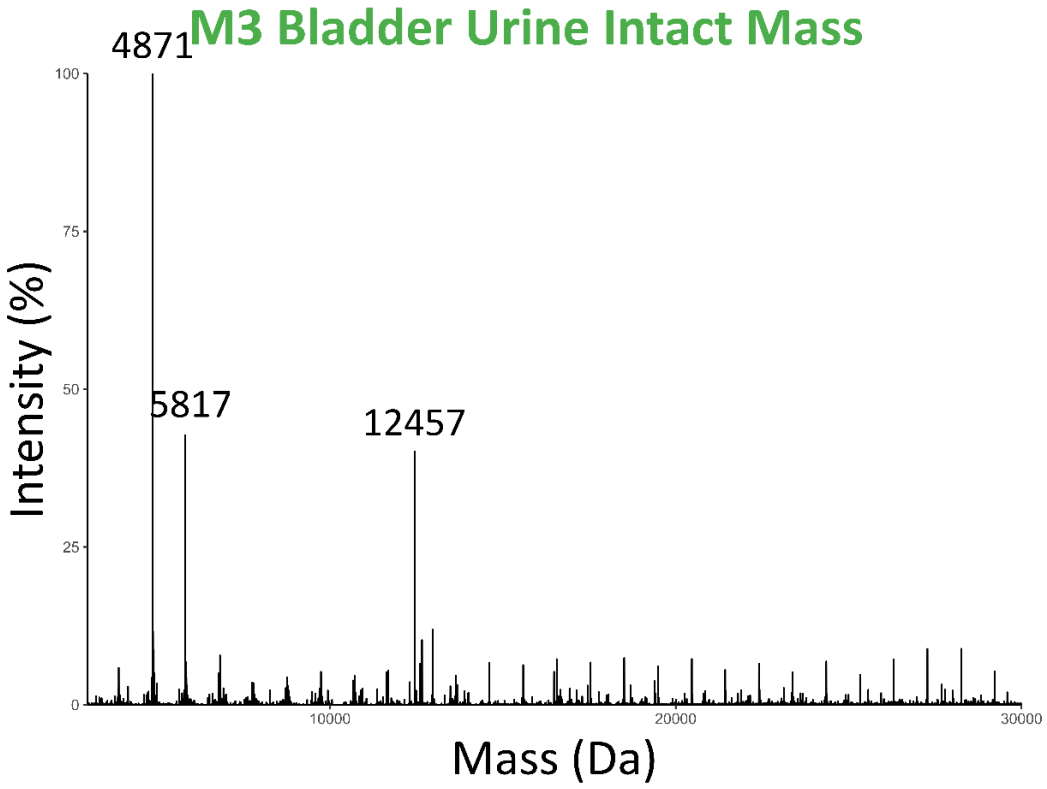


Figure 3.51 Intact mass of bank vole bladder urine from male 3.

3.6.7 Female scent mark intact protein mass spectra

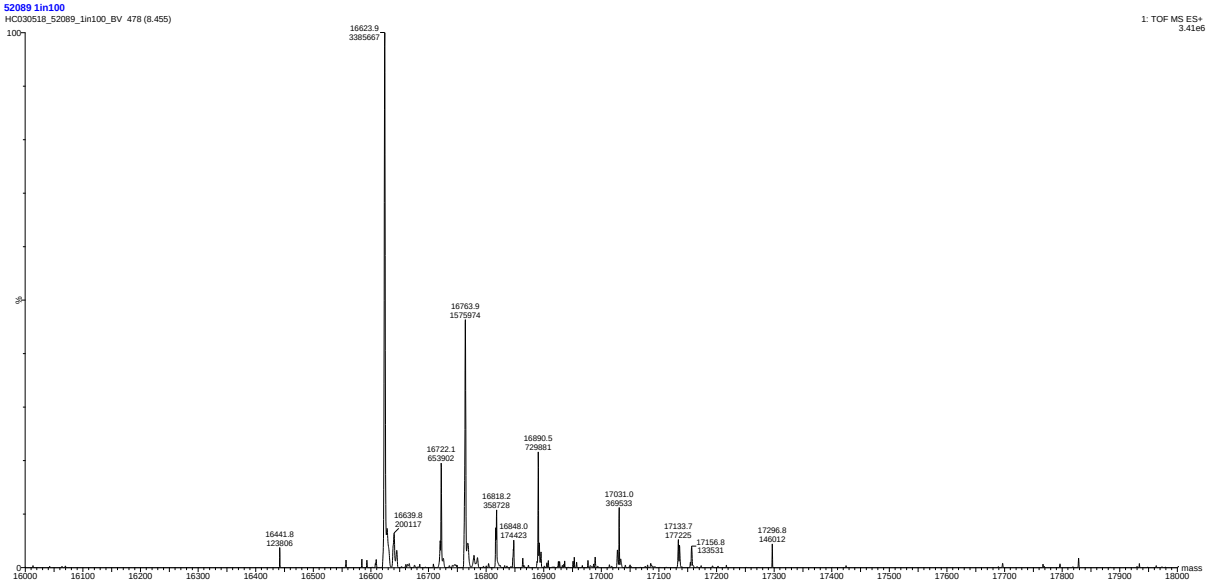


Figure 3.52 Intact protein profile of bank vole scent marks from female 2. Intact mass analysis performed by Holly Coombes (Mammalian Behaviour & Evolution, University of Liverpool).

4 Characterisation of the urinary protein content in the field vole, *Microtus agrestis*

4.6 Supplementary Material

4.6.1 Intact mass profiles of urine from mature field voles

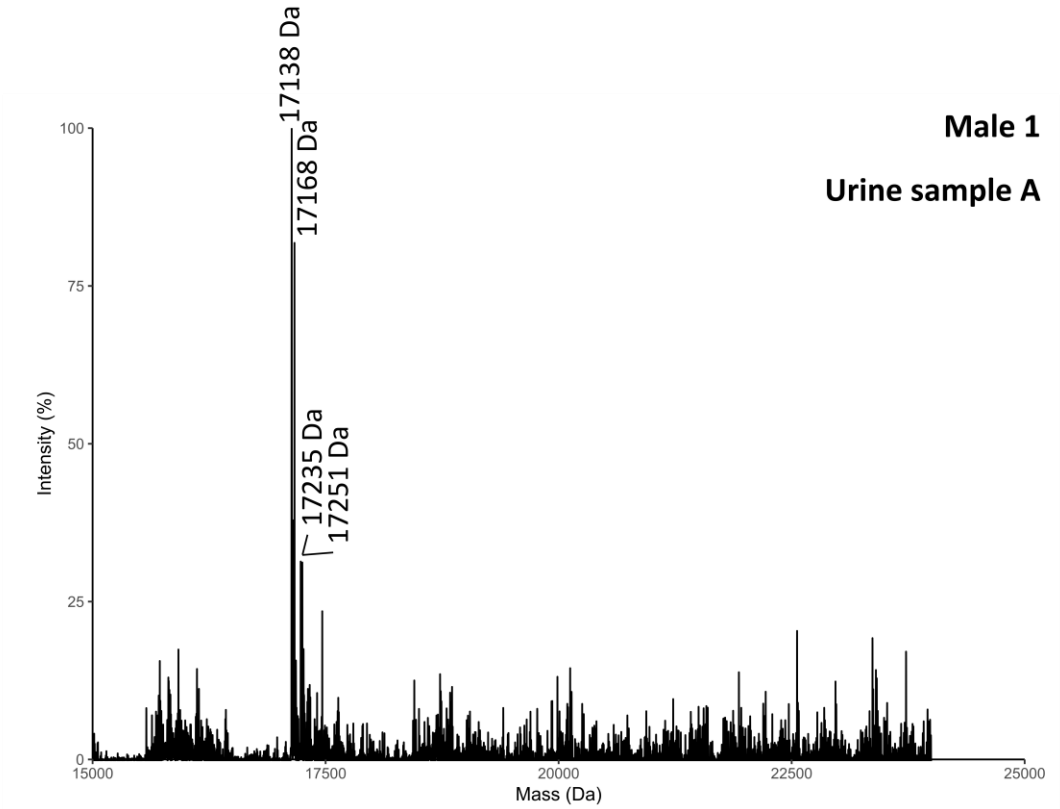


Figure 4.1 LC-MS analysis of intact protein from urine of captive male field vole 1, sample A.

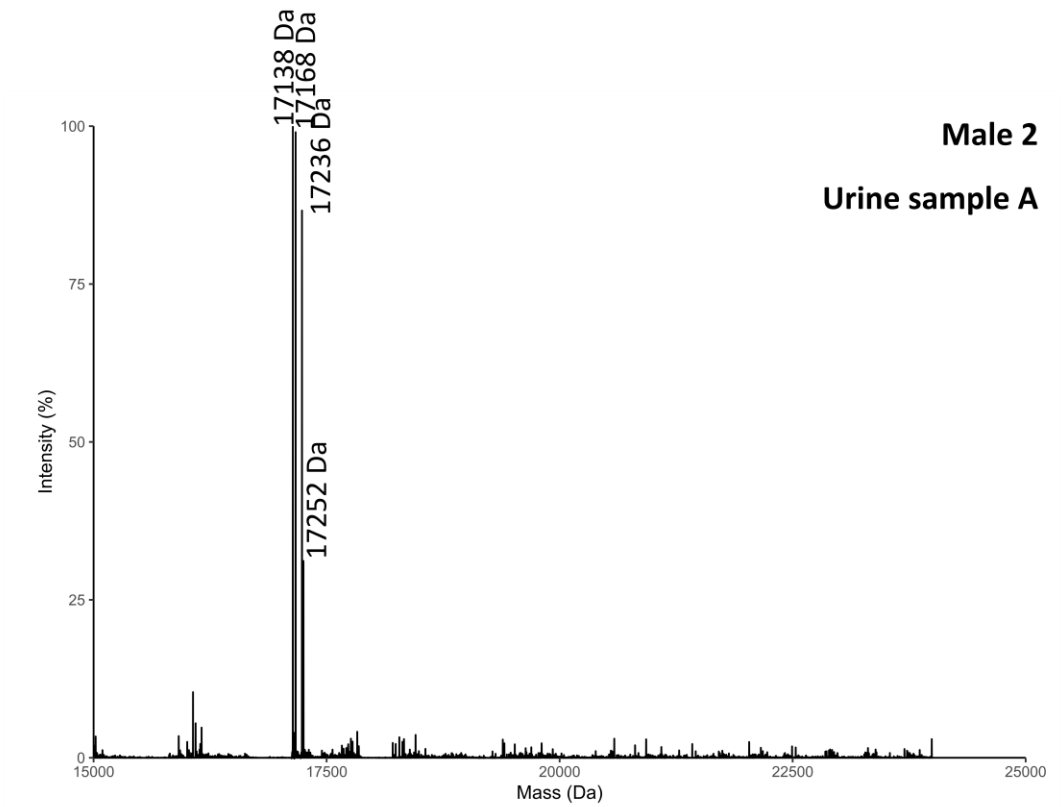


Figure 4.2 LC-MS analysis of intact protein from urine of captive male field vole 2, sample A.

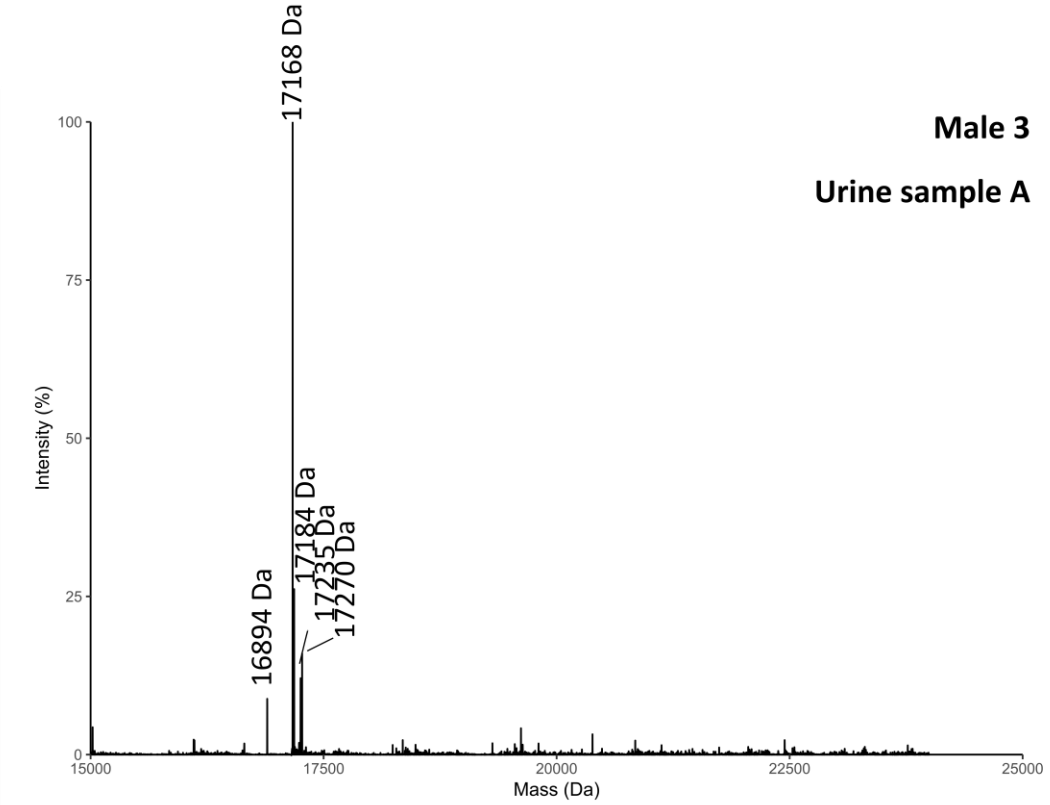


Figure 4.3 LC-MS analysis of intact protein from urine of captive male field vole 3, sample A.

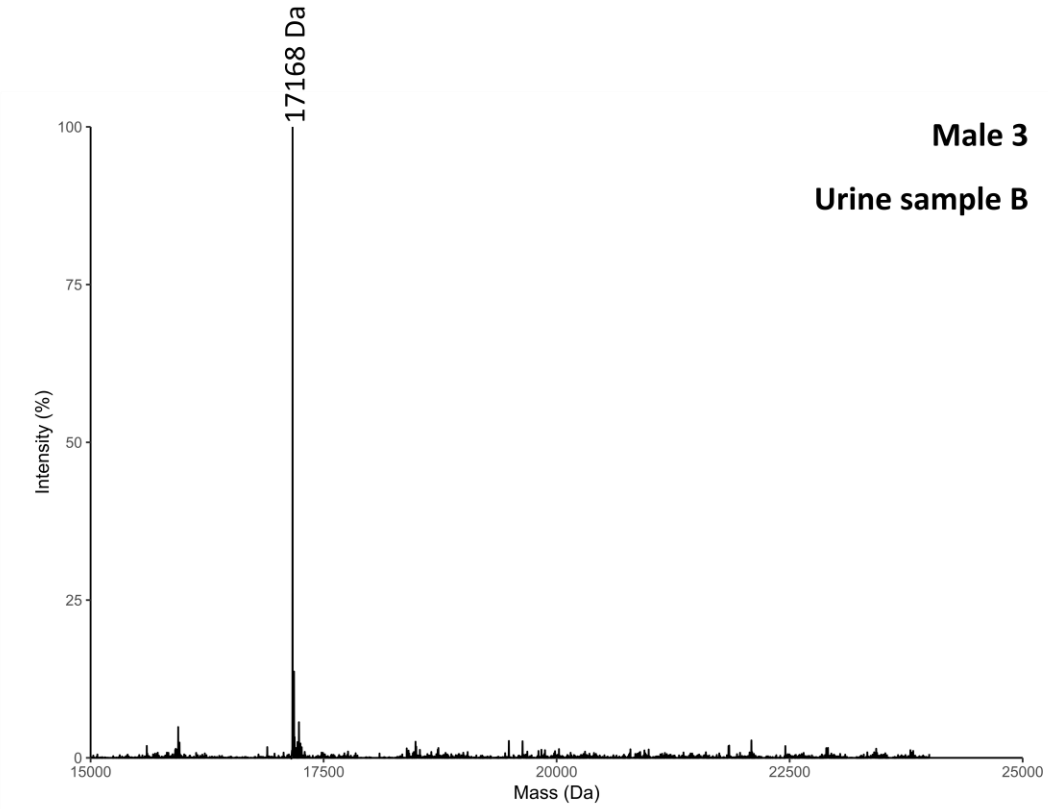


Figure 4.4 LC-MS analysis of intact protein from urine of captive male field vole 3, sample B.

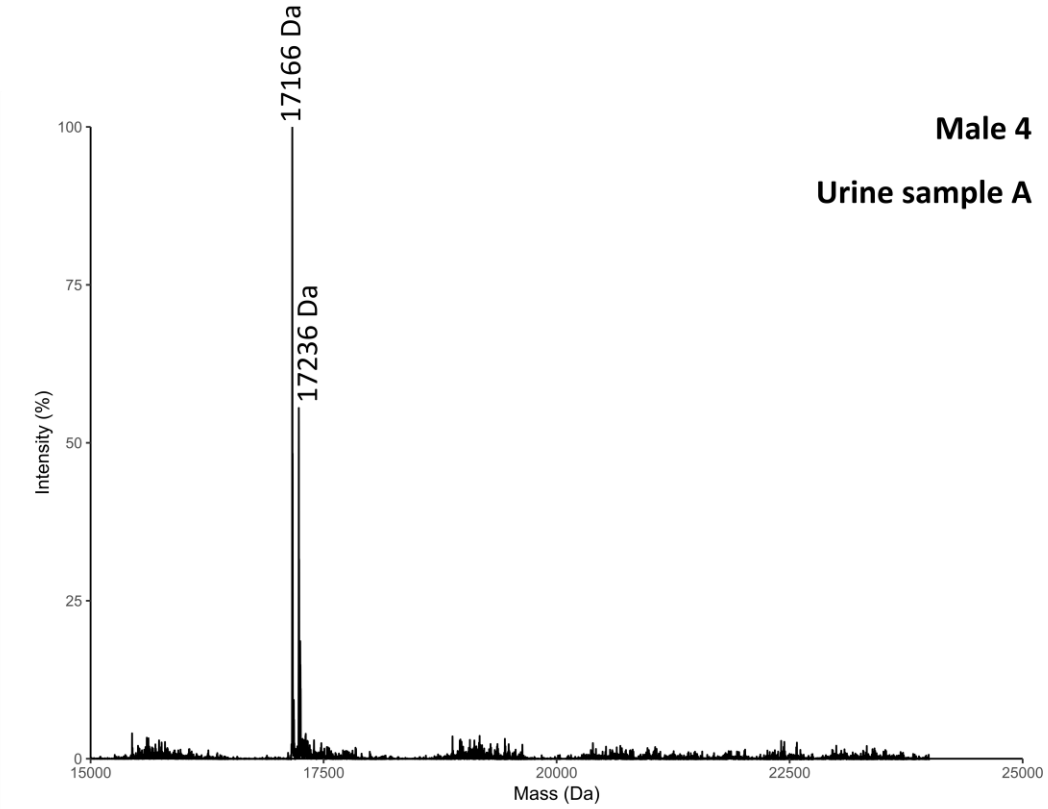


Figure 4.5 LC-MS analysis of intact protein from urine of captive male field vole 4, sample A.

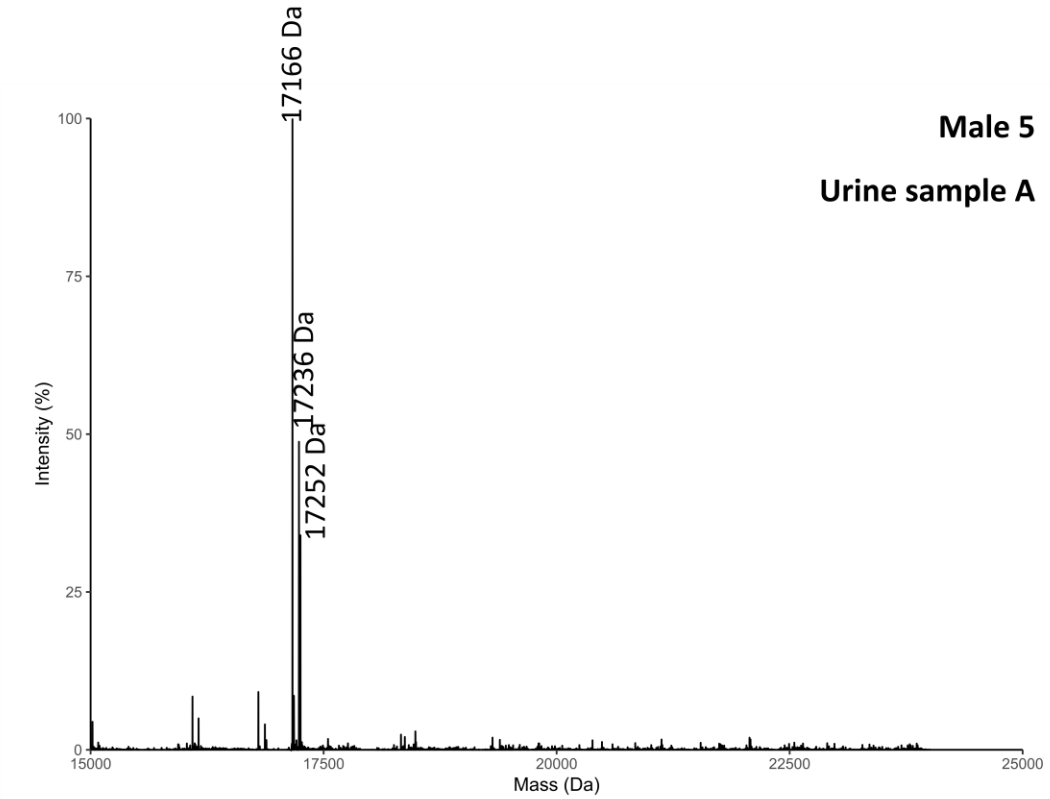


Figure 4.6 LC-MS analysis of intact protein from urine of captive male field vole 5, sample A.

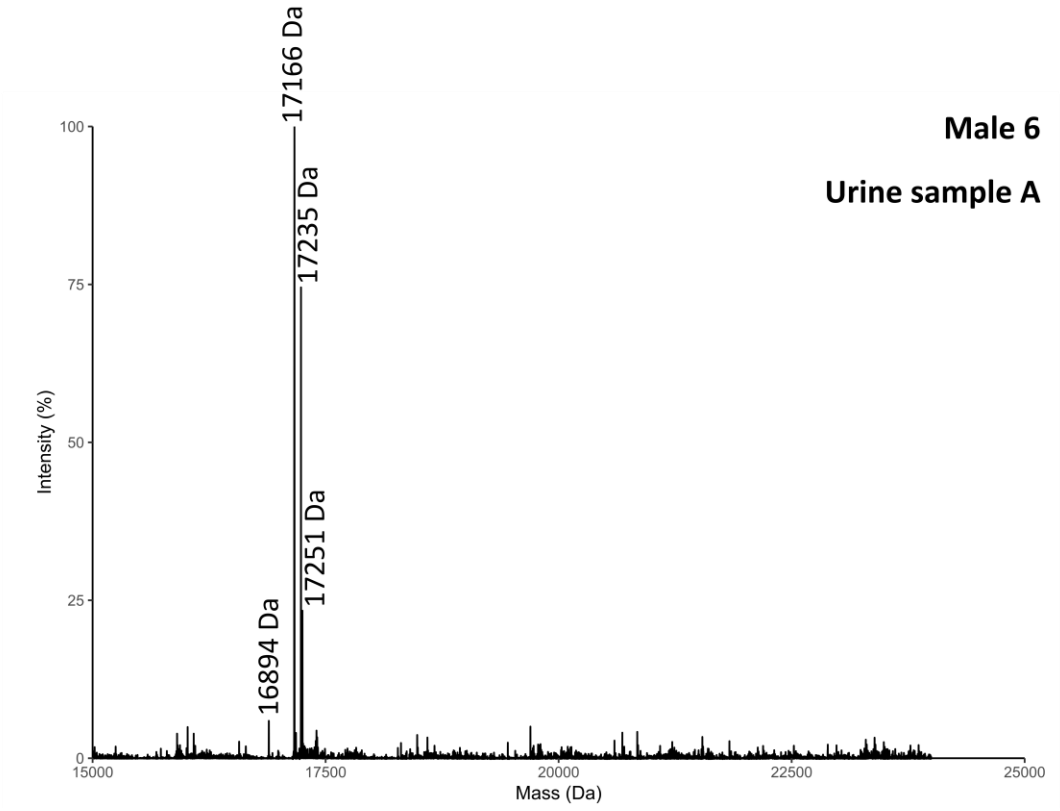


Figure 4.7 LC-MS analysis of intact protein from urine of captive male field vole 6, sample A.

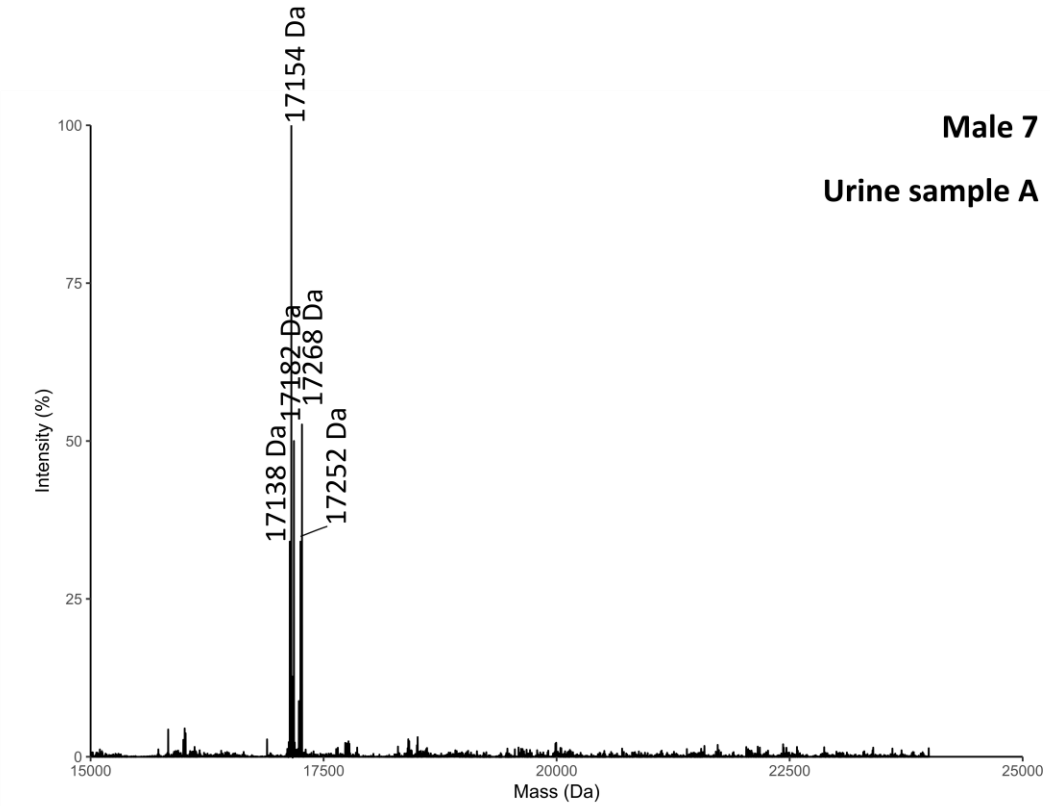


Figure 4.8 LC-MS analysis of intact protein from urine of captive male field vole 7, sample A.

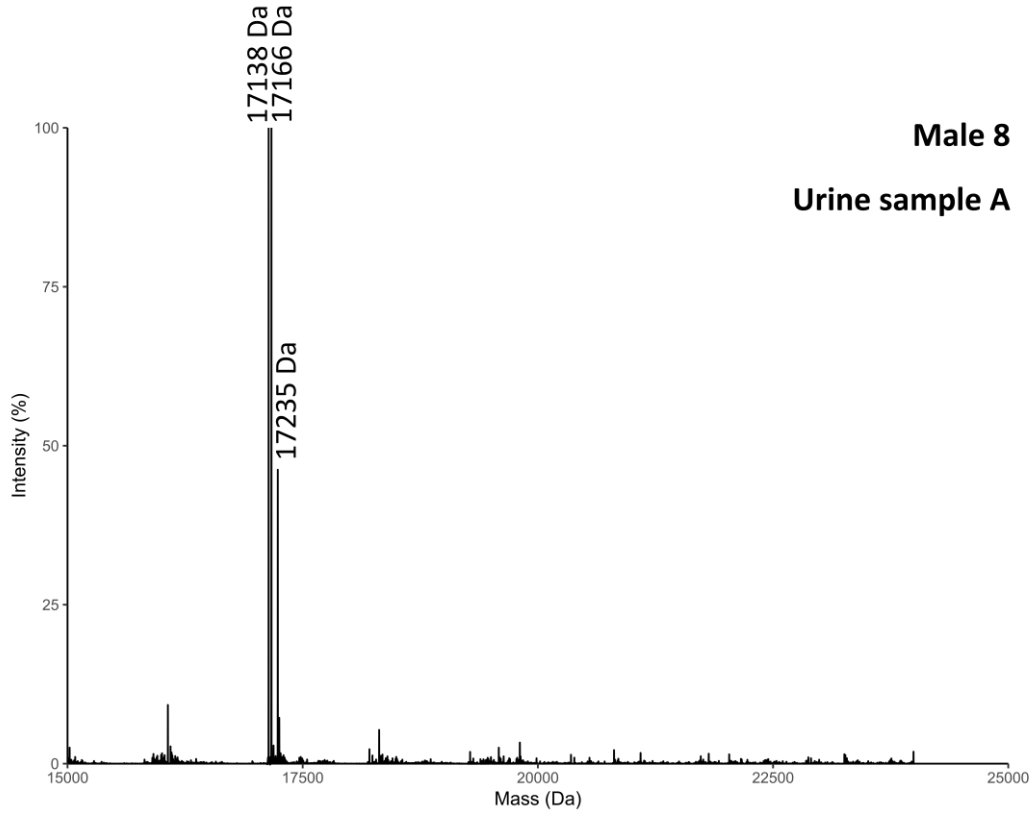


Figure 4.9 LC-MS analysis of intact protein from urine of captive male field vole 8, sample A.

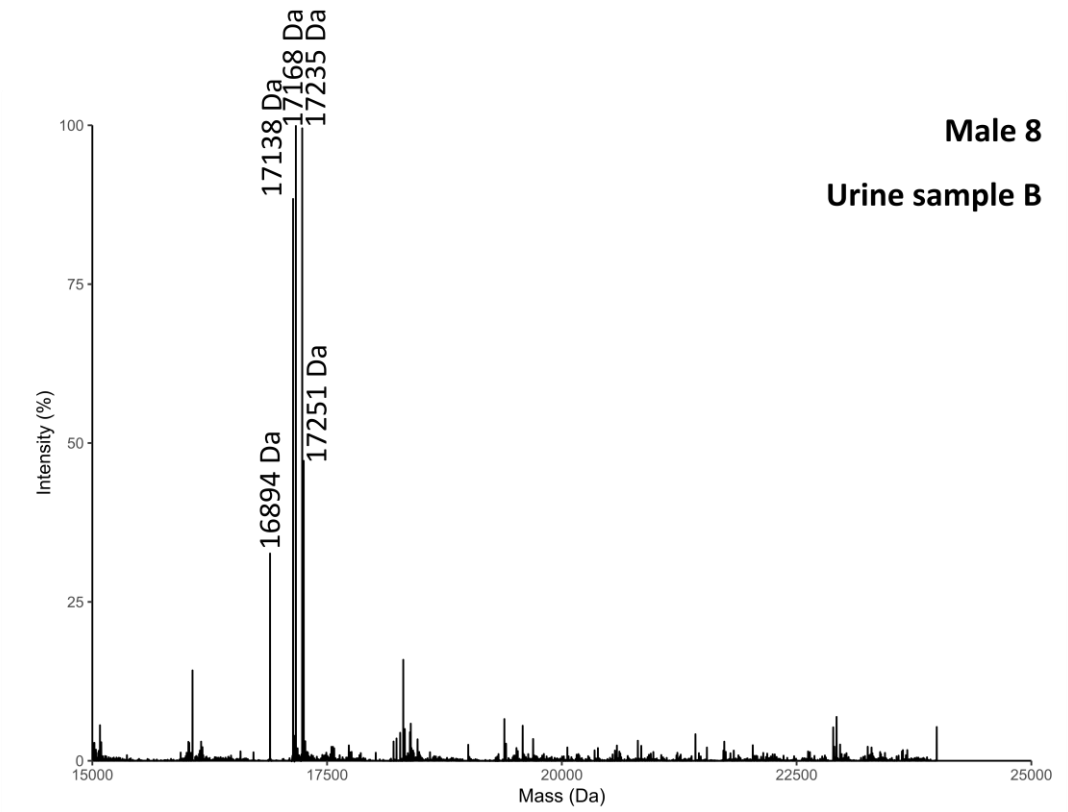


Figure 4.10 LC-MS analysis of intact protein from urine of captive male field vole 8, sample B.

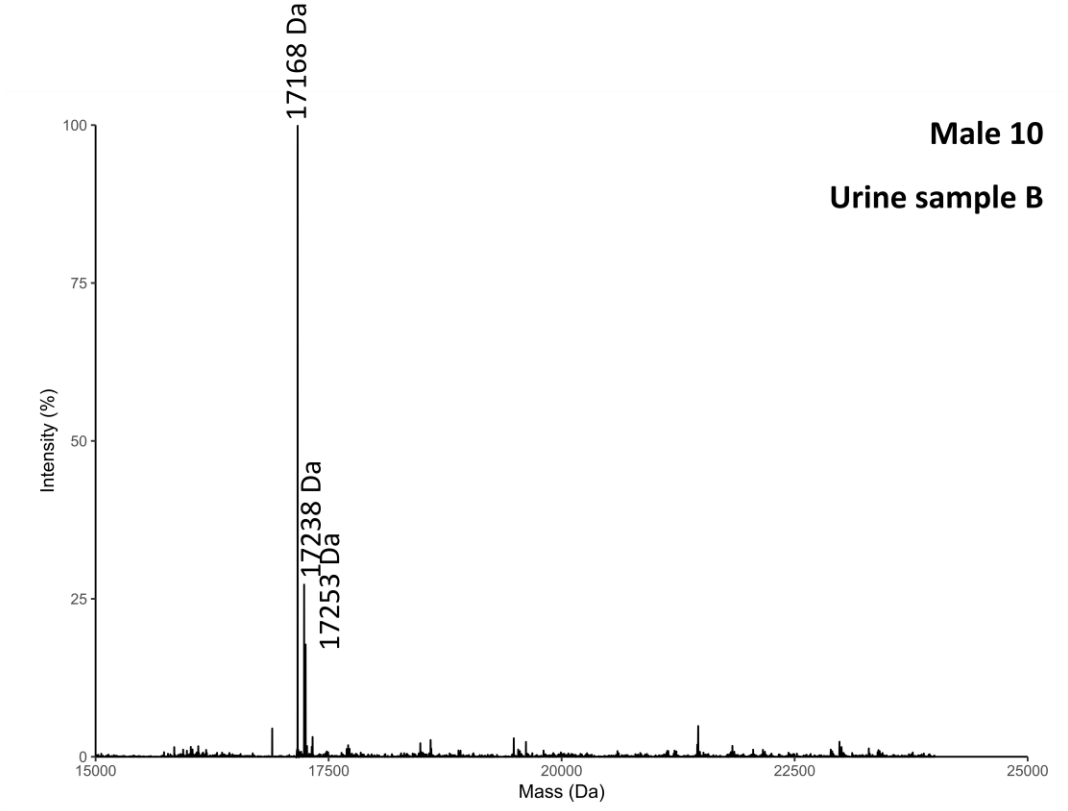


Figure 4.11 LC-MS analysis of intact protein from urine of captive male field vole 10, sample B.

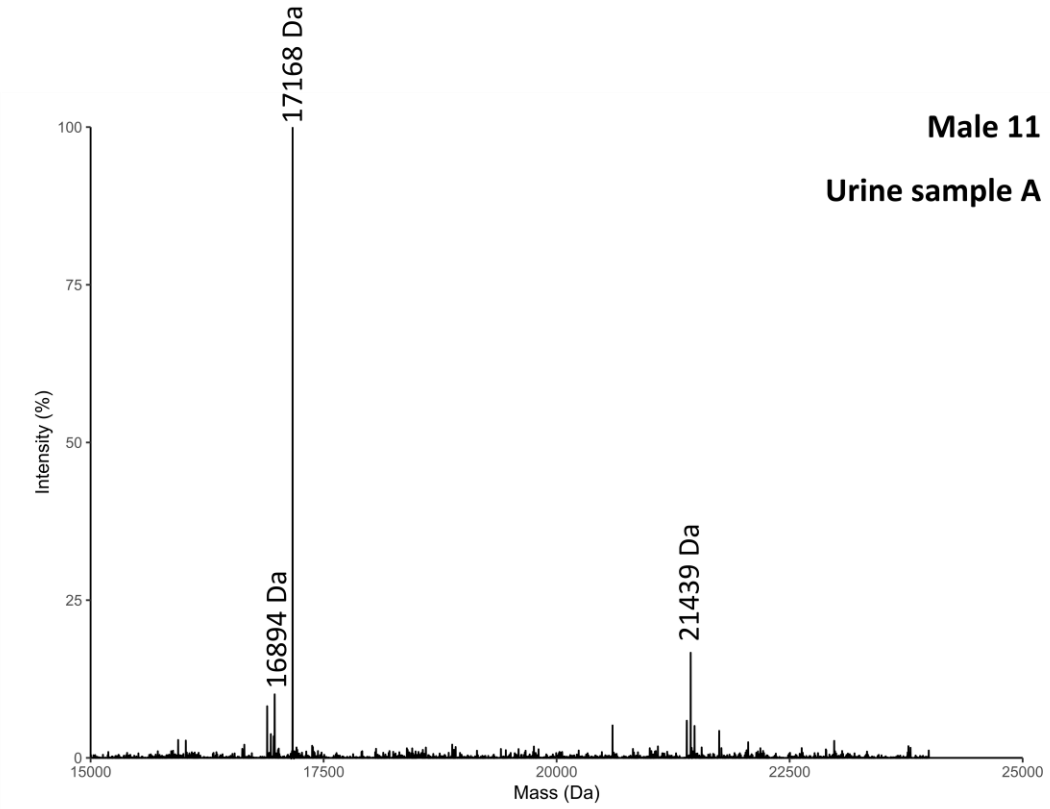


Figure 4.12 LC-MS analysis of intact protein from urine of captive male field vole 11, sample A.

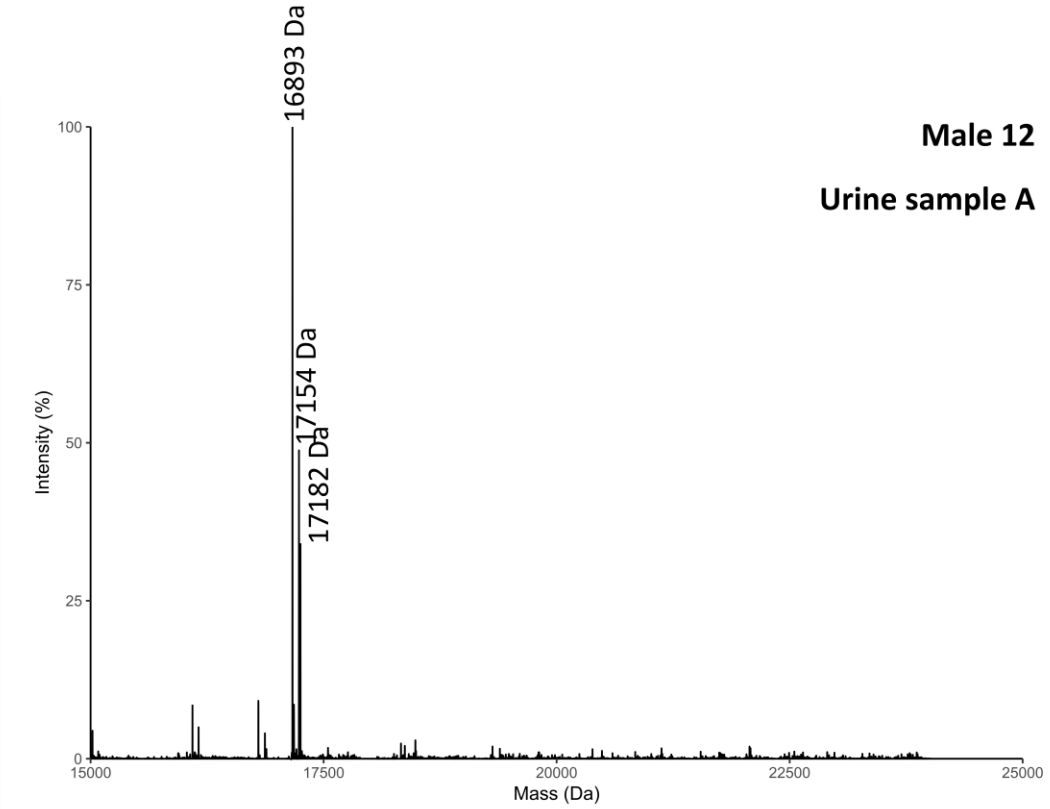


Figure 4.13 LC-MS analysis of intact protein from urine of captive male field vole 12, sample A.

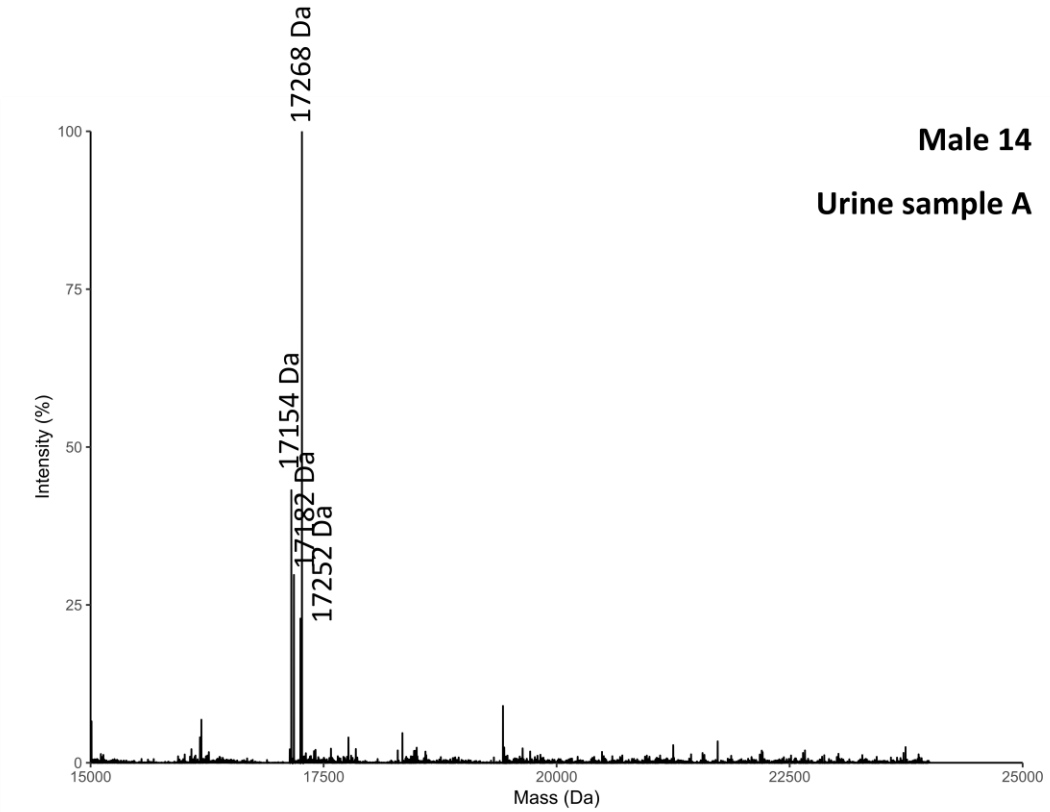


Figure 4.14 LC-MS analysis of intact protein from urine of captive male field vole 14, sample A.

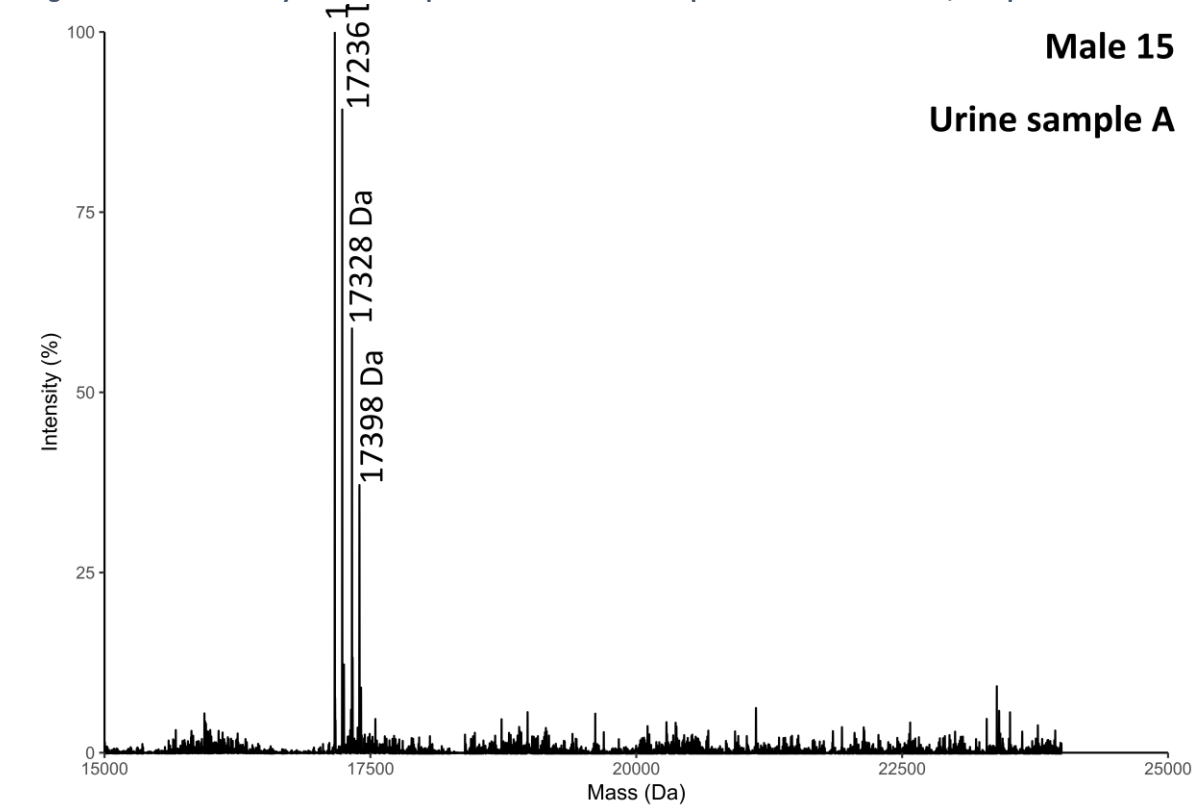
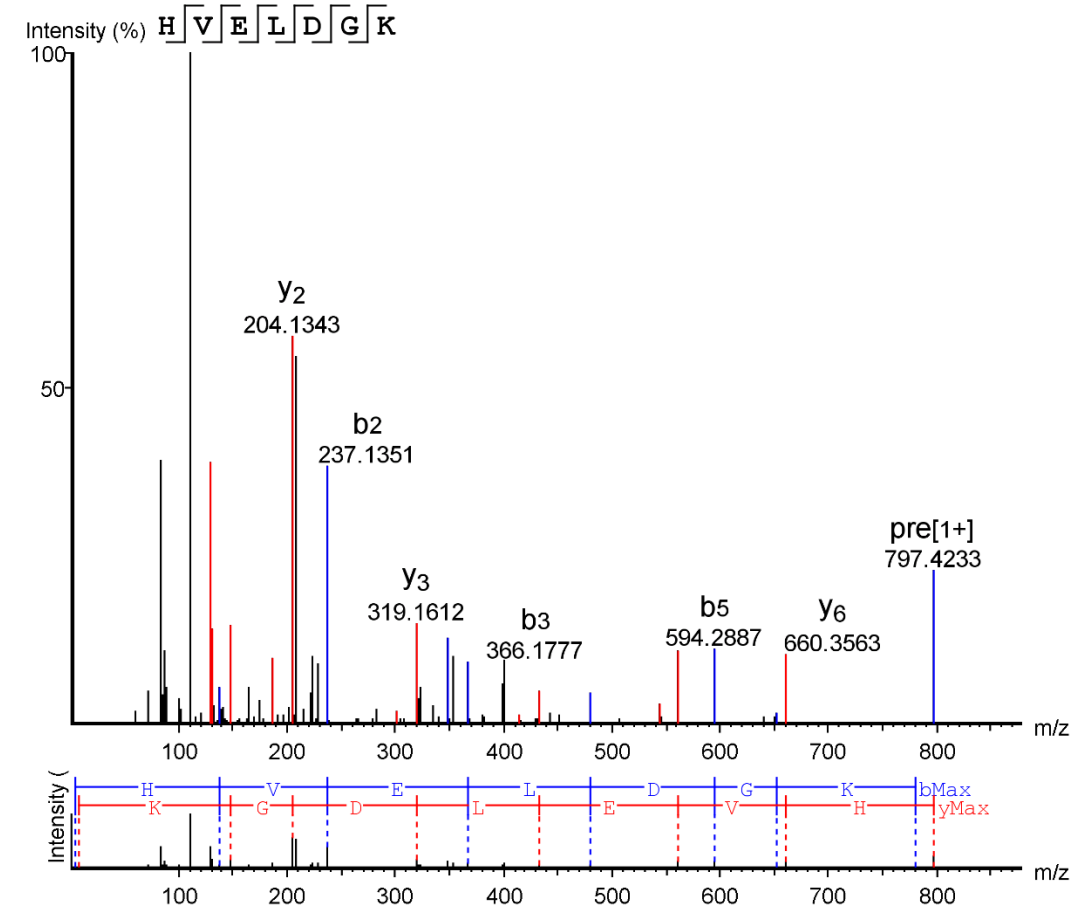


Figure 4.15 LC-MS analysis of intact protein from urine of captive male field vole 15, sample A.

4.6.2 *Fragment ion spectra of peptides used to sequence field vole glareosin.*

4.6.2.1 *Initial sequence*



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	138.0662	120.0562	121.0397	69.5334	H					7
2	237.1342	219.1246	220.1081	119.0676	V	660.3558	642.3457	643.3293	330.6781	6
3	366.1768	348.1667	349.1508	183.5889	E	561.2869	543.2772	544.2609	281.1439	5
4	479.2607	461.2512	462.2348	240.1309	L	432.2449	414.2329	415.2183	216.6226	4
5	594.2870	576.2781	577.2617	297.6443	D	319.1611	301.1503	302.1342	160.0806	3
6	651.3058	633.2996	634.2832	326.1551	G	204.1343	186.1236	187.1074	102.5671	2
7					K	147.1127	129.1023	130.0863	74.0564	1

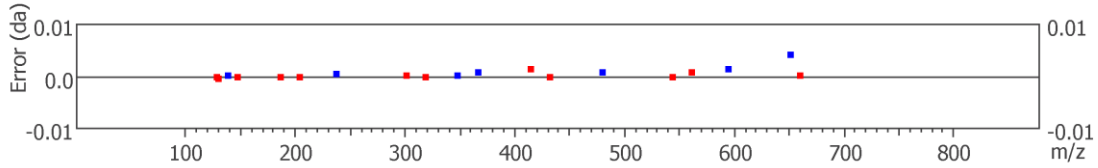
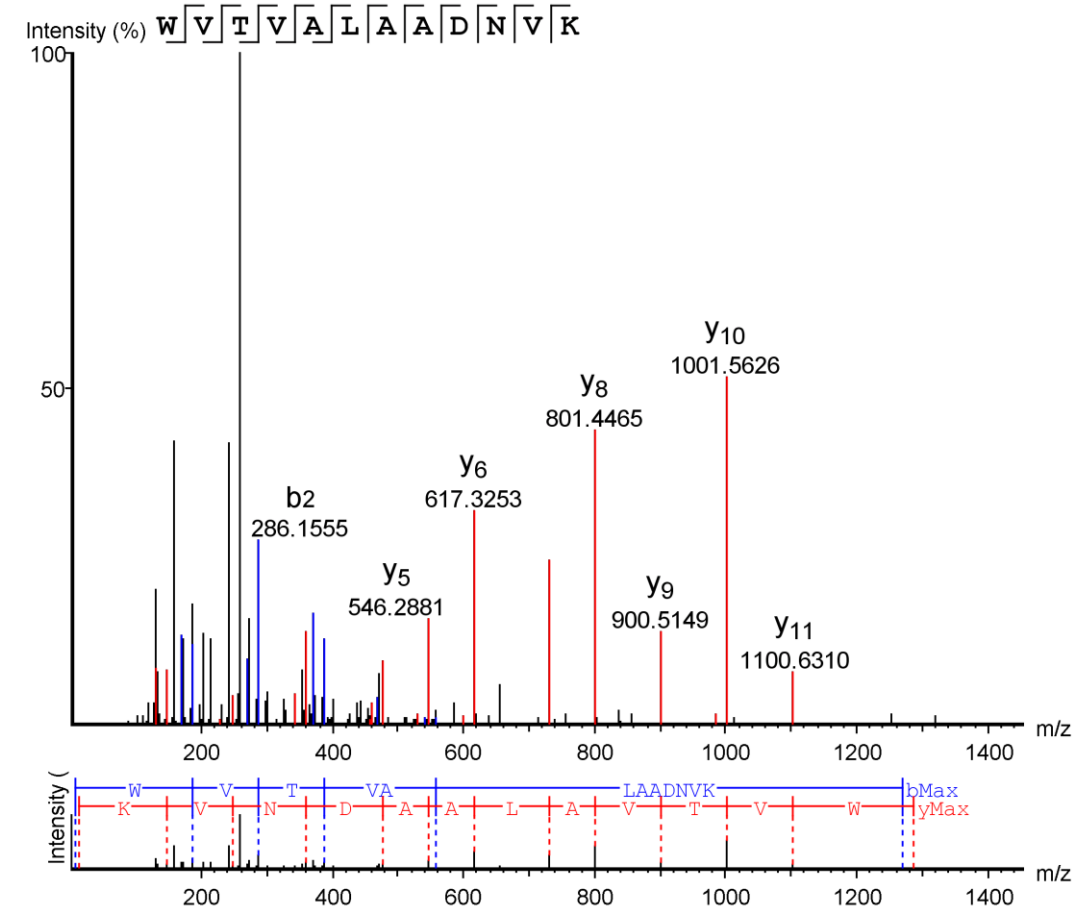


Figure 4.16 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t1.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	187.0867	169.0766	170.0602	94.0436	W					12
2	286.1549	268.1450	269.1284	143.5778	V	1100.6284	1082.6205	1083.6040	550.8155	11
3	387.2027	369.1921	370.1752	194.1016	T	1001.5619	983.5515	984.5454	501.2813	10
4	486.2716	468.2601	469.2447	243.6358	V	900.5126	882.5043	883.4879	450.7574	9
5	557.3091	539.2930	540.2828	279.1544	A	801.4460	783.4359	784.4195	401.2232	8
6	670.3928	652.3823	653.3658	335.6964	L	730.4088	712.3988	713.3824	365.7047	7
7	741.4299	723.4194	724.4030	371.2150	A	617.3249	599.3147	600.2985	309.1626	6
8	812.4670	794.4565	795.4401	406.7335	A	546.2883	528.2777	529.2614	273.6441	5
9	927.4940	909.4834	910.4670	464.2470	D	475.2511	457.2393	458.2252	238.1255	4
10	1041.5369	1023.5264	1024.5099	521.2684	N	360.2239	342.2136	343.1970	180.6121	3
11	1140.6053	1122.5948	1123.5784	570.8027	V	246.1811	228.1709	229.1542	123.5906	2
12					K	147.1128	129.1024	130.0862	74.0564	1

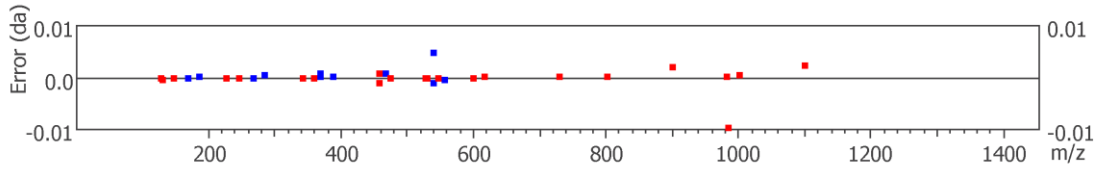
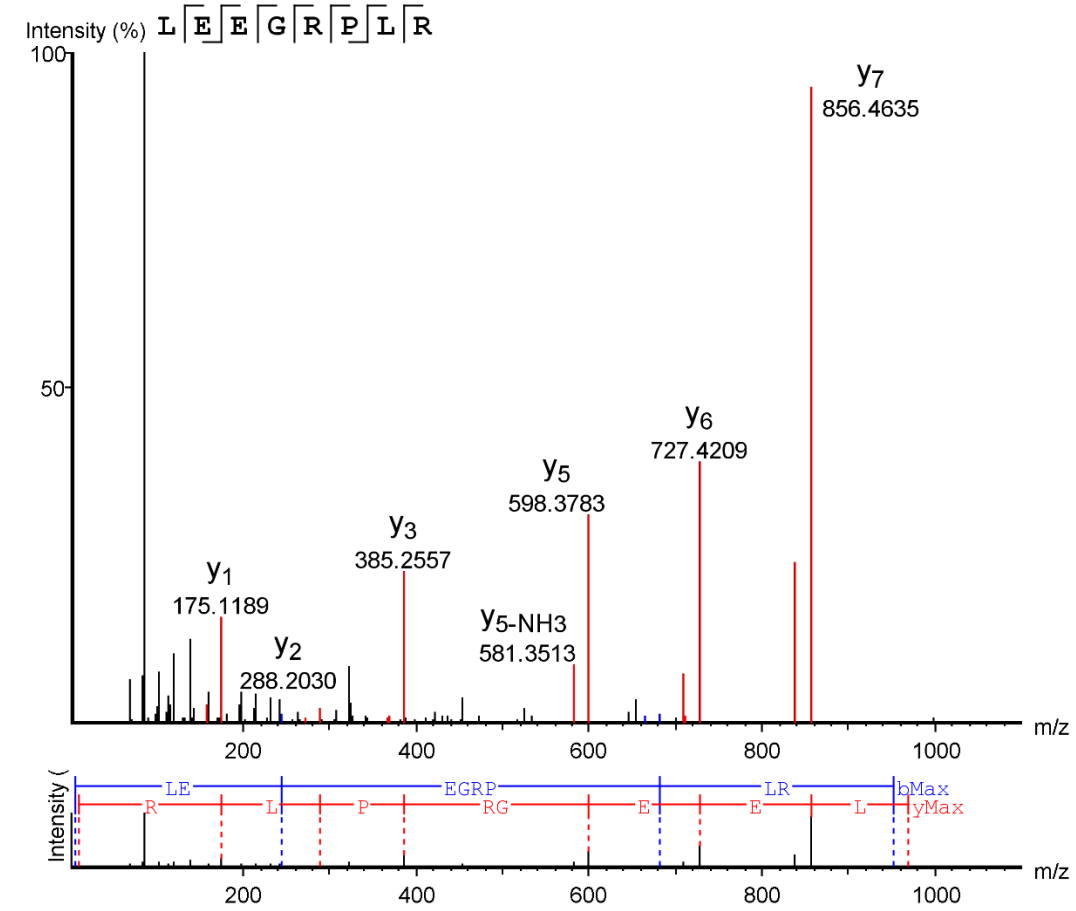


Figure 4.17 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t2.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					8
2	243.1340	225.1239	226.1075	122.0672	E	856.4629	838.4528	839.4365	428.7318	7
3	372.1771	354.1665	355.1501	186.5885	E	727.4203	709.4082	710.3981	364.2104	6
4	429.1985	411.1880	412.1715	215.0993	G	598.3778	580.3677	581.3516	299.6891	5
5	585.2996	567.2891	568.2726	293.1498	R	541.3569	523.3463	524.3299	271.1783	4
6	682.3526	664.3373	665.3254	341.6762	P	385.2553	367.2452	368.2300	193.1279	3
7	795.4364	777.4258	778.4094	398.2182	L	288.2031	270.1924	271.1783	144.6015	2
8					R	175.1190	157.1084	158.0924	88.0595	1

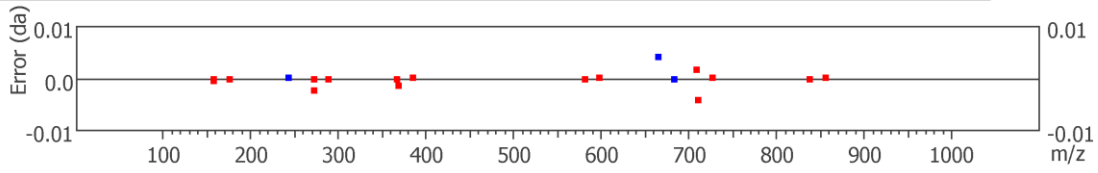
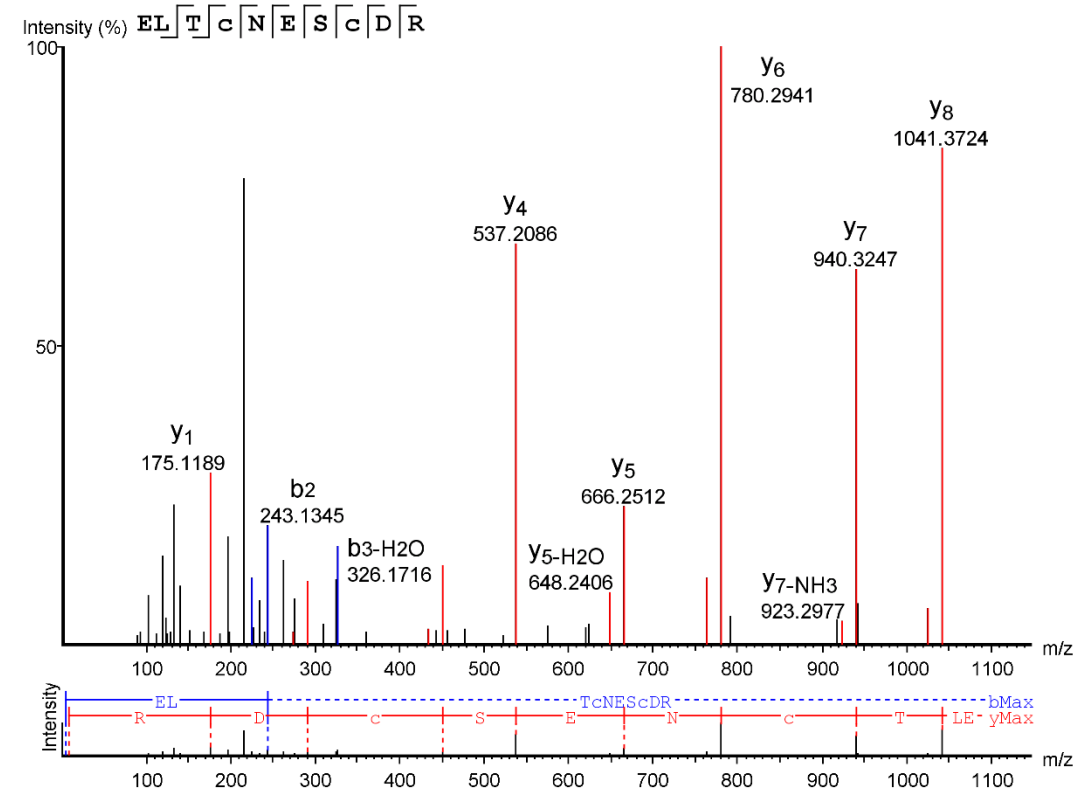


Figure 4.18 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t3.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	130.0504	112.0399	113.0234	65.5252	E					10
2	243.1337	225.1233	226.1075	122.0672	L	1154.4564	1136.4459	1137.4294	577.7282	9
3	344.1821	326.1705	327.1552	172.5911	T	1041.3711	1023.3618	1024.3511	521.1862	8
4	504.2128	486.2022	487.1858	252.6064	C(+57.02)	940.3260	922.3141	923.2933	470.6624	7
5	618.2557	600.2452	601.2288	309.6279	N	780.2925	762.2835	763.2675	390.6470	6
6	747.2983	729.2878	730.2714	374.1492	E	666.2488	648.2377	649.2242	333.6256	5
7	834.3303	816.3198	817.3033	417.6652	S	537.2082	519.1980	520.1816	269.1043	4
8	994.3610	976.3505	977.3340	497.6805	C(+57.02)	450.1763	432.1660	433.1519	225.5883	3
9	1109.3879	1091.3774	1092.3610	555.1940	D	290.1459	272.1353	273.1195	145.5729	2
10					R	175.1189	157.1084	158.0919	88.0595	1

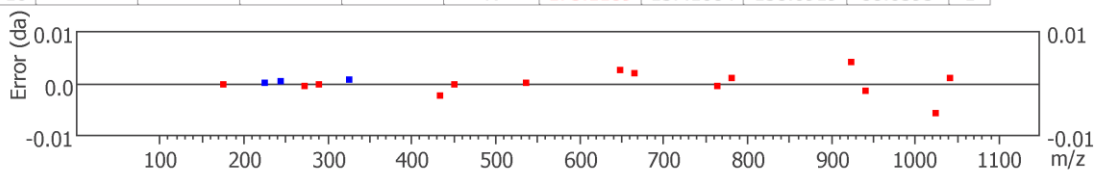
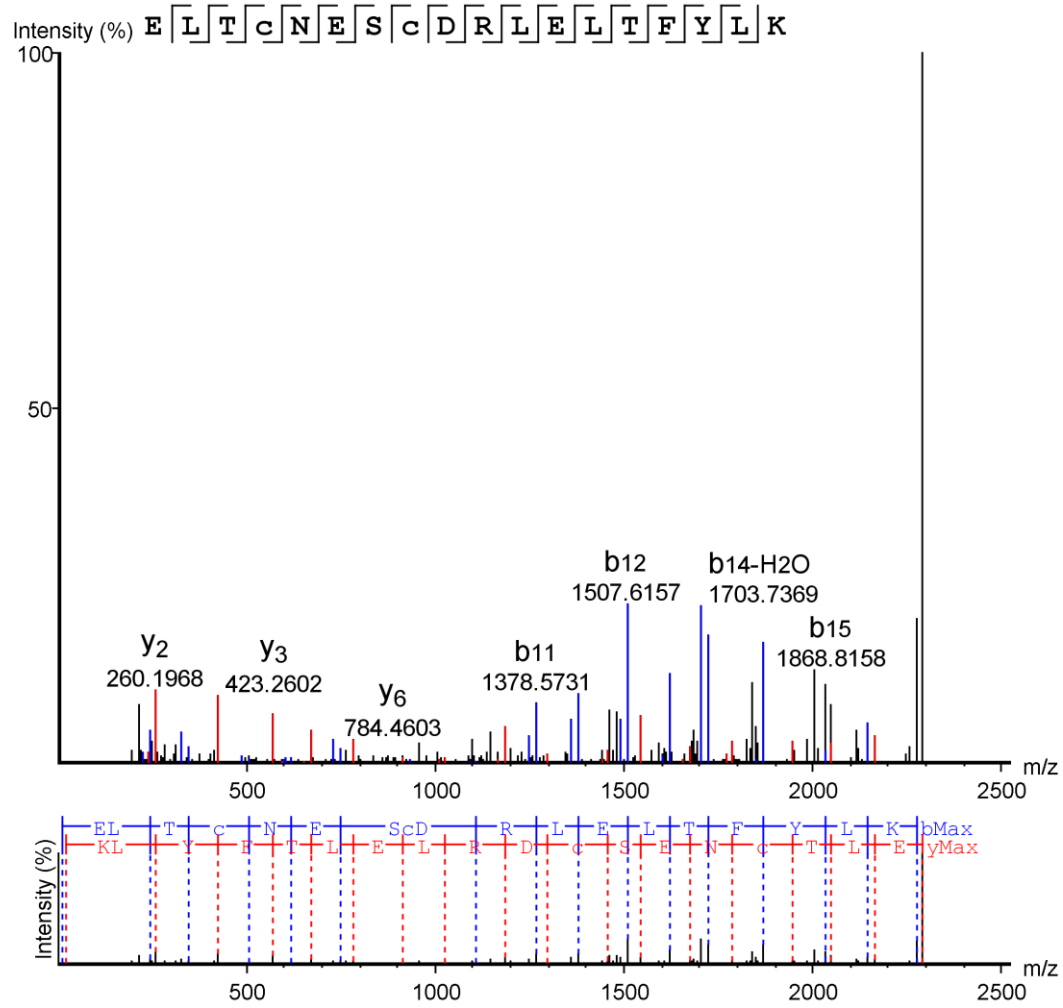


Figure 4.19 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t4.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	130.0504	112.0399	113.0234	65.5252	E					18
2	243.1338	225.1231	226.1075	122.0672	L	2162.0237	2144.0151	2144.9985	1081.5128	17
3	344.1800	326.1707	327.1552	172.5911	T	2048.9424	2030.9310	2031.9146	1024.9708	16
4	504.2121	486.2007	487.1858	252.6064	C(+57.02)	1947.8876	1929.8833	1930.8668	974.4469	15
5	618.2543	600.2458	601.2291	309.6279	N	1787.8624	1769.8527	1770.8406	894.4316	14
6	747.2950	729.2866	730.2714	374.1492	E	1673.8176	1655.8196	1656.7933	837.4102	13
7	834.3303	816.3198	817.3033	417.6652	S	1544.7780	1526.7672	1527.7507	772.8889	12
8	994.3610	976.3505	977.3340	497.6805	C(+57.02)	1457.7405	1439.7351	1440.7146	729.3728	11
9	1109.3843	1091.3755	1092.3610	555.1940	D	1297.7094	1279.7045	1280.6880	649.3575	10
10	1265.4917	1247.4795	1248.4620	633.2445	R	1182.6868	1164.6776	1165.6682	591.8458	9
11	1378.5752	1360.5649	1361.5461	689.7866	L	1026.5883	1008.5809	1009.5600	513.7935	8
12	1507.6155	1489.5985	1490.5889	754.3079	E	913.5004	895.4924	896.4760	457.2515	7
13	1620.6964	1602.6906	1603.6727	810.8499	L	784.4604	766.4498	767.4333	392.7302	6
14	1721.7463	1703.7350	1704.7205	861.3737	T	671.3760	653.3627	654.3492	336.1881	5
15	1868.8149	1850.7961	1851.7888	934.9086	F	570.3286	552.3180	553.3016	285.6643	4
16	2031.8761	2013.8687	2014.8522	1016.4396	Y	423.2599	405.2496	406.2332	212.1301	3
17	2144.9590	2126.9526	2127.9360	1072.9816	L	260.1966	242.1861	243.1701	130.5984	2
18					K	147.1128	129.1022	130.0858	74.0564	1

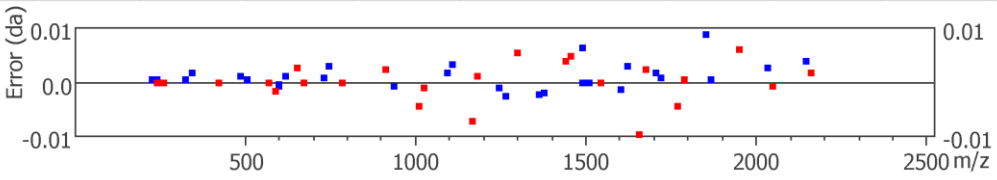
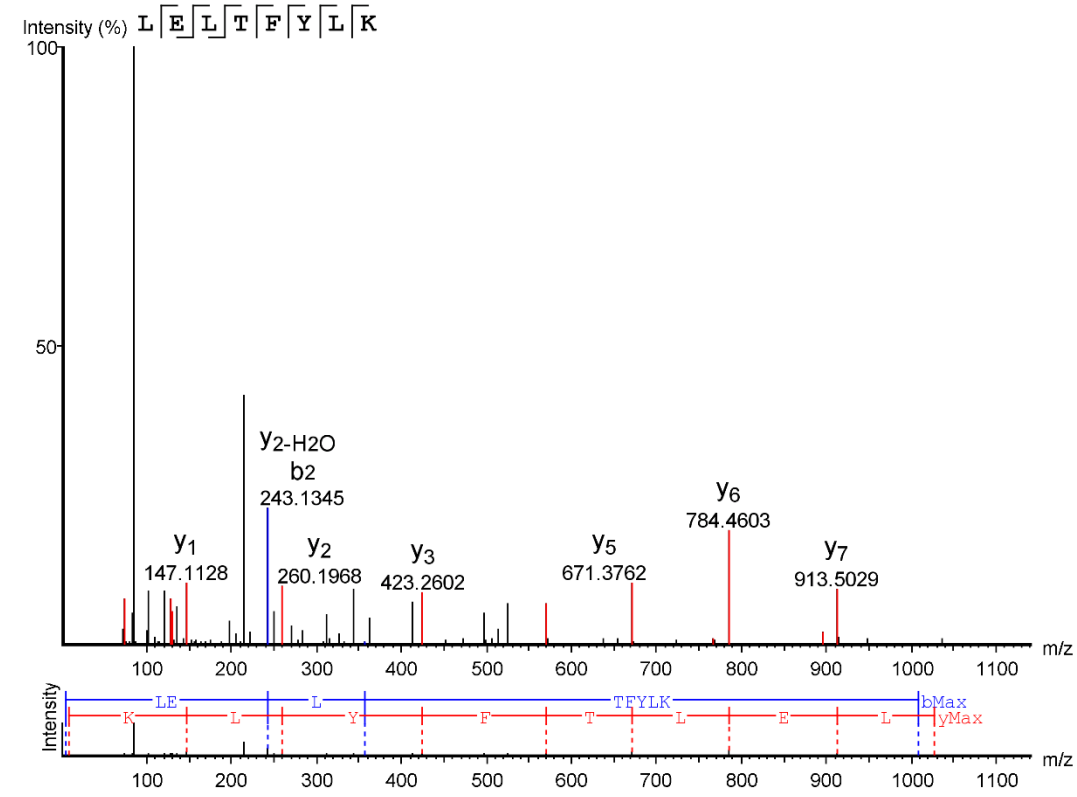


Figure 4.20 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t5.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					8
2	243.1338	225.1239	226.1075	122.0672	E	913.4977	895.4918	896.4760	457.2515	7
3	356.2185	338.2079	339.1915	178.6093	L	784.4583	766.4525	767.4333	392.7302	6
4	457.2662	439.2556	440.2392	229.1331	T	671.3759	653.3657	654.3492	336.1881	5
5	604.3346	586.3240	587.3076	302.6673	F	570.3297	552.3180	553.3016	285.6643	4
6	767.3979	749.3874	750.3710	384.1990	Y	423.2593	405.2496	406.2332	212.1301	3
7	880.4820	862.4714	863.4550	440.7410	L	260.1966	242.1862	243.1698	130.5984	2
8					K	147.1126	129.1022	130.0862	74.0607	1

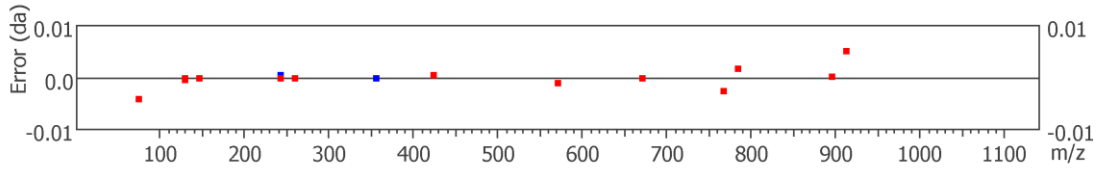
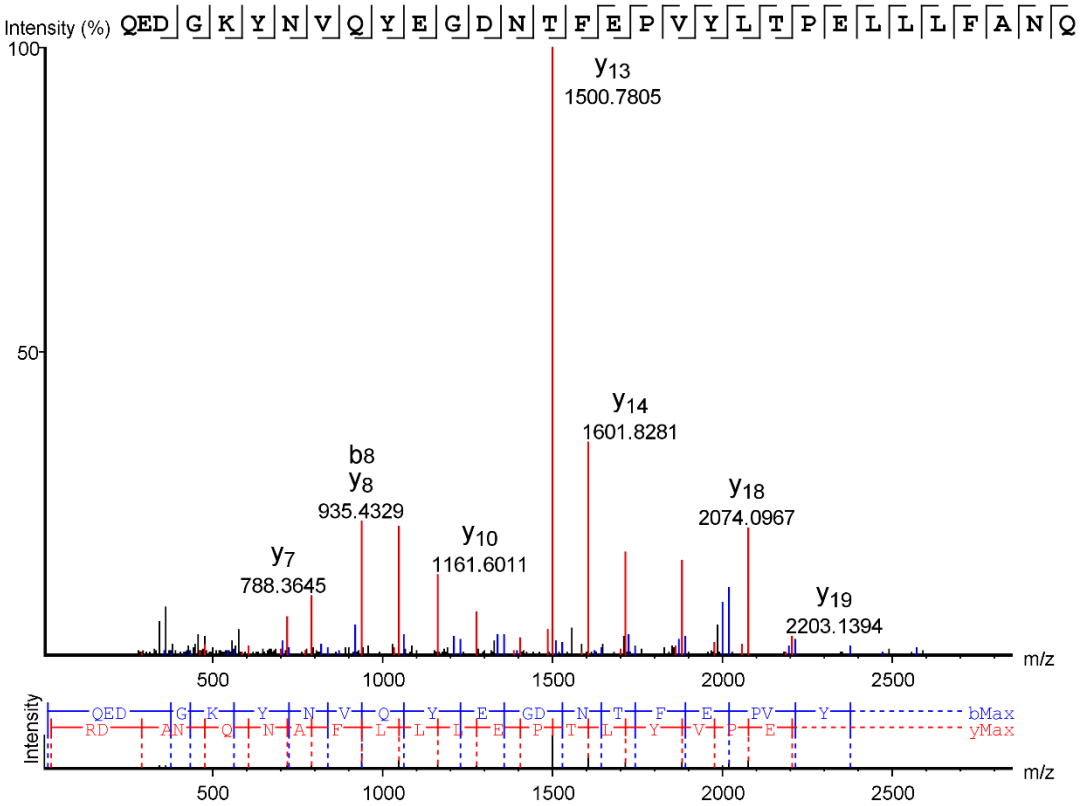


Figure 4.21 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t6.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	129.0664	111.0558	112.0394	65.0332	Q					35
2	258.1090	240.0984	241.0820	129.5545	E	3962.8718	3944.8613	3945.8447	1981.9359	34
3	373.1348	355.1245	356.1082	187.0680	D	3833.8291	3815.8186	3816.8020	1917.4146	33
4	430.1555	412.1456	413.1281	215.5787	G	3718.8022	3700.7917	3701.7751	1859.9011	32
5	558.2523	540.2412	541.2238	279.6262	K	3661.7808	3643.7703	3644.7537	1831.3904	31
6	721.3149	703.3036	704.2875	361.1578	Y	3533.6858	3515.6753	3516.6587	1767.3429	30
7	835.3553	817.3462	818.3306	418.1793	N	3370.6226	3352.6121	3353.5955	1685.8113	29
8	934.4249	916.4142	917.4006	467.7135	V	3256.5796	3238.5691	3239.5525	1628.7898	28
9	1062.4833	1044.4720	1045.4601	531.7428	Q	3157.5112	3139.5007	3140.4841	1579.2556	27
10	1225.5464	1207.5381	1208.5229	613.2745	Y	3029.4526	3011.4421	3012.4255	1515.2263	26
11	1354.5902	1336.5786	1337.5667	677.7958	E	2866.3894	2848.3789	2849.3623	1433.6947	25
12	1411.6129	1393.6119	1394.5811	706.3065	G	2737.3467	2719.3362	2720.3196	1369.1733	24
13	1526.6333	1508.6294	1509.6101	763.8199	D	2680.3252	2662.3147	2663.2981	1340.6626	23
14	1640.6804	1622.6724	1623.6593	820.8414	N	2565.2983	2547.2878	2548.2712	1283.1492	22
15	1741.7311	1723.7202	1724.7101	871.3647	T	2451.2554	2433.2449	2434.2283	1226.1277	21
16	1888.7913	1870.7885	1871.7749	944.8995	F	2350.2078	2332.1973	2333.1807	1175.6039	20
17	2017.8370	1999.8311	2000.8190	1009.4208	E	2203.1311	2185.1372	2186.1123	1102.0697	19
18	2114.8943	2096.8838	2097.8672	1057.9471	P	2074.0957	2056.0825	2057.0696	1037.5483	18
19	2213.9561	2195.9470	2196.9385	1107.4813	V	1977.0437	1959.0334	1960.0170	989.0220	17
20	2377.0210	2359.0154	2359.9988	1189.0129	Y	1877.9740	1859.9700	1860.9486	939.4878	16
21	2490.1101	2472.0955	2473.0830	1245.5551	L	1714.9117	1696.9100	1697.8853	857.9561	15
22	2591.1577	2573.1455	2574.1306	1296.0789	T	1601.8263	1583.8237	1584.8011	801.4141	14
23	2688.2104	2670.2000	2671.1833	1344.6052	P	1500.7780	1482.7732	1483.7598	750.8903	13
24	2817.2532	2799.2427	2800.2261	1409.1266	E	1403.7267	1385.7151	1386.7007	702.3638	12
25	2930.3372	2912.3267	2913.3101	1465.6686	L	1274.6854	1256.6746	1257.6581	637.8425	11
26	3043.4211	3025.4106	3026.3940	1522.2106	L	1161.5997	1143.5906	1144.5741	581.3005	10
27	3156.5054	3138.4949	3139.4783	1578.7527	L	1048.5150	1030.5065	1031.4873	524.7585	9
28	3303.5737	3285.5632	3286.5466	1652.2869	F	935.4324	917.4224	918.4030	468.2165	8
29	3374.6108	3356.6003	3357.5837	1687.8054	A	788.3630	770.3539	771.3368	394.6823	7
30	3488.6538	3470.6433	3471.6267	1744.8269	N	717.3260	699.3168	700.3005	359.1637	6
31	3616.7124	3598.7019	3599.6853	1808.8562	Q	603.2838	585.2739	586.2541	302.1422	5
32	3730.7551	3712.7446	3713.7280	1865.8776	N	475.2252	457.2153	458.1980	238.1129	4
33	3801.7922	3783.7817	3784.7651	1901.3961	A	361.1830	343.1724	344.1560	181.0915	3
34	3916.8193	3898.8088	3899.7922	1958.9097	D	290.1457	272.1353	273.1189	145.5729	2
35					R	175.1189	157.1084	158.0919	88.0595	1

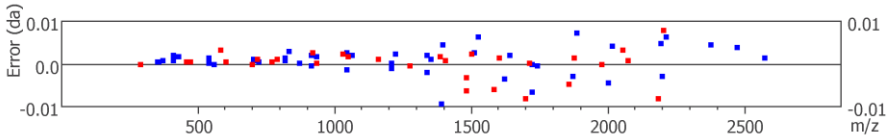
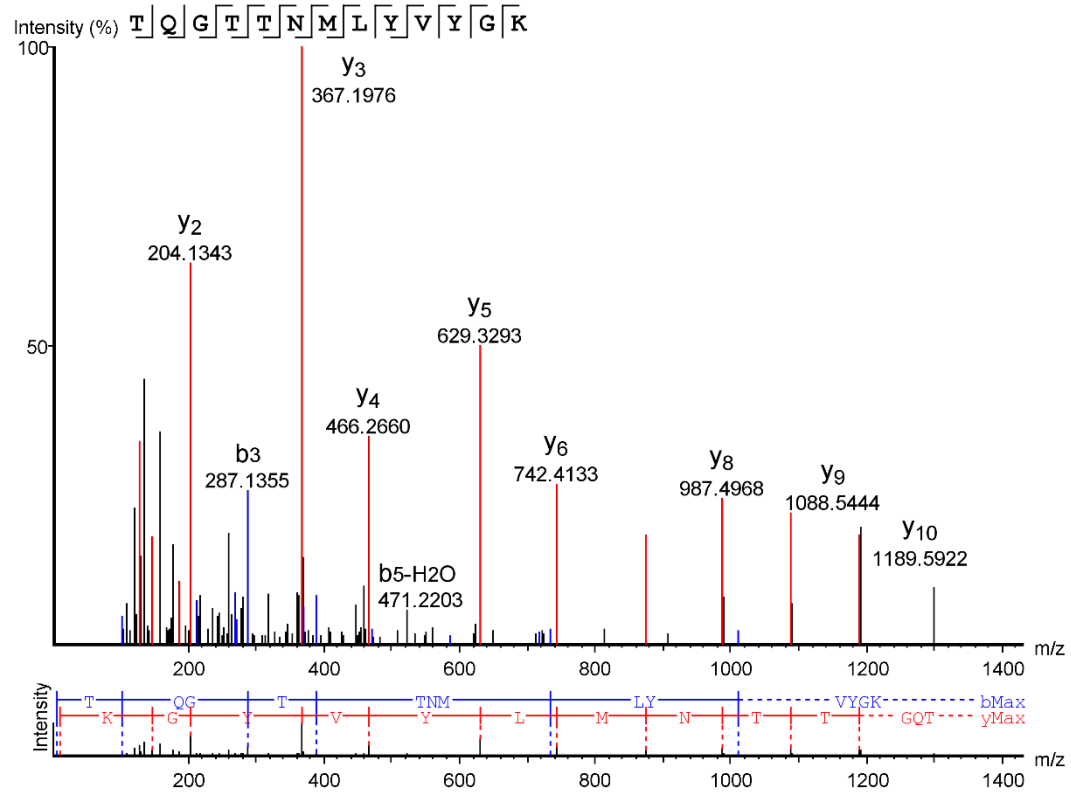


Figure 4.22 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t7.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	102.0551	84.0450	85.0285	51.5278	T					13
2	230.1141	212.1026	213.0865	115.5570	Q	1374.6722	1356.6617	1357.6453	687.8361	12
3	287.1346	269.1238	270.1079	144.0678	G	1246.6136	1228.6031	1229.5867	623.8068	11
4	388.1816	370.1711	371.1562	194.5916	T	1189.5939	1171.5817	1172.5652	595.2961	10
5	489.2309	471.2183	472.2039	245.1154	T	1088.5367	1070.5339	1071.5175	544.7722	9
6	603.2738	585.2632	586.2430	302.1369	N	987.4951	969.4862	970.4698	494.2484	8
7	734.3125	716.3038	717.2874	367.6572	M	873.4501	855.4433	856.4269	437.2269	7
8	847.3984	829.3878	830.3714	424.1992	L	742.4100	724.4028	725.3864	371.7067	6
9	1010.4628	992.4512	993.4348	505.7309	Y	629.3282	611.3187	612.3023	315.1646	5
10	1109.5302	1091.5197	1092.5032	555.2651	V	466.2659	448.2554	449.2390	233.6330	4
11	1272.5935	1254.5830	1255.5665	636.7968	Y	367.1973	349.1870	350.1706	184.0988	3
12	1329.6149	1311.6044	1312.5879	665.3074	G	204.1342	186.1235	187.1076	102.5671	2
13					K	147.1127	129.1024	130.0858	74.0564	1

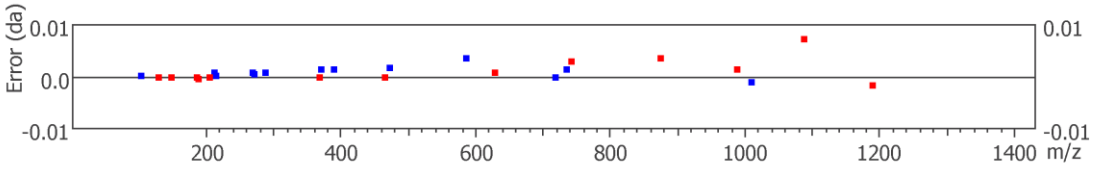
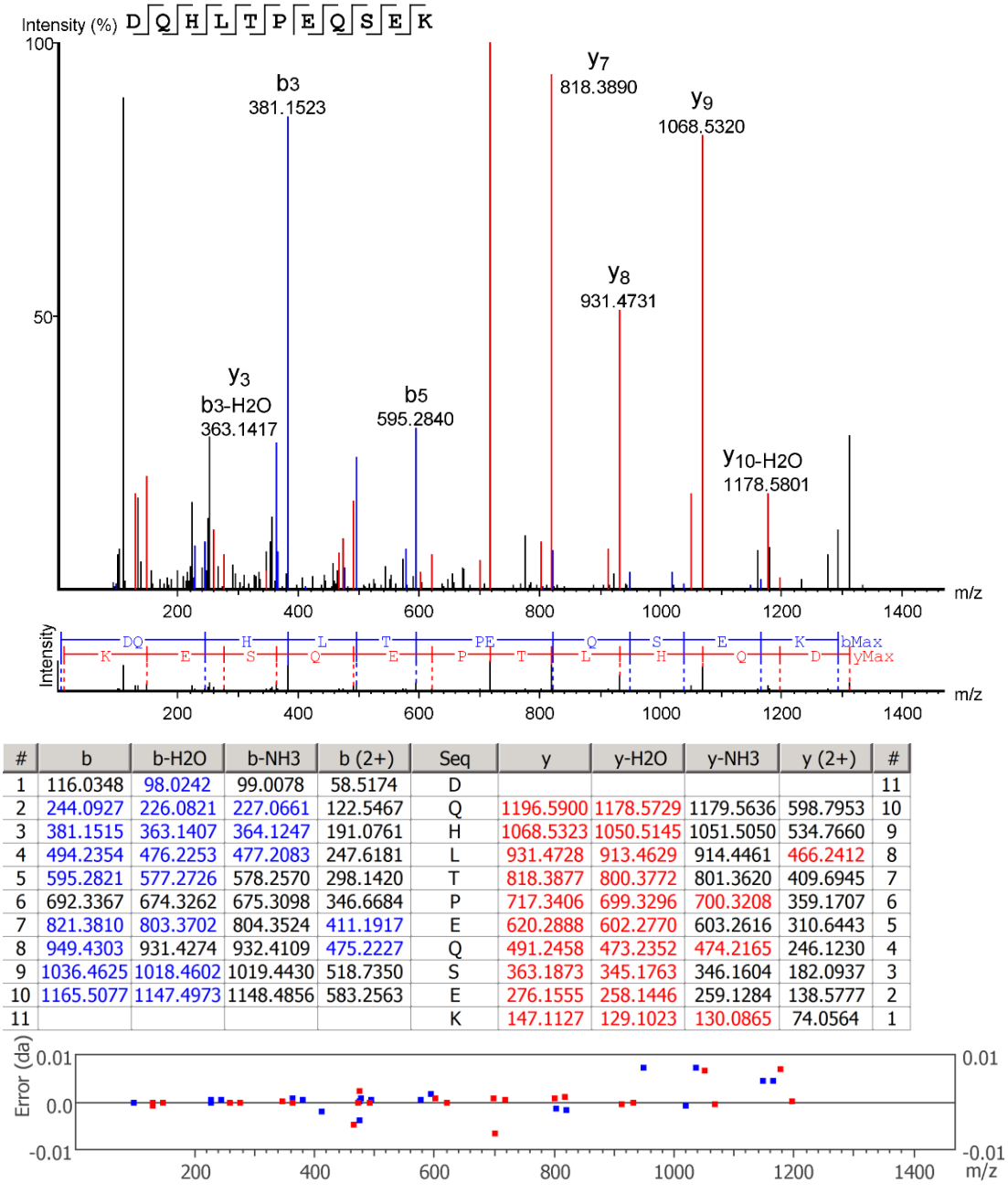
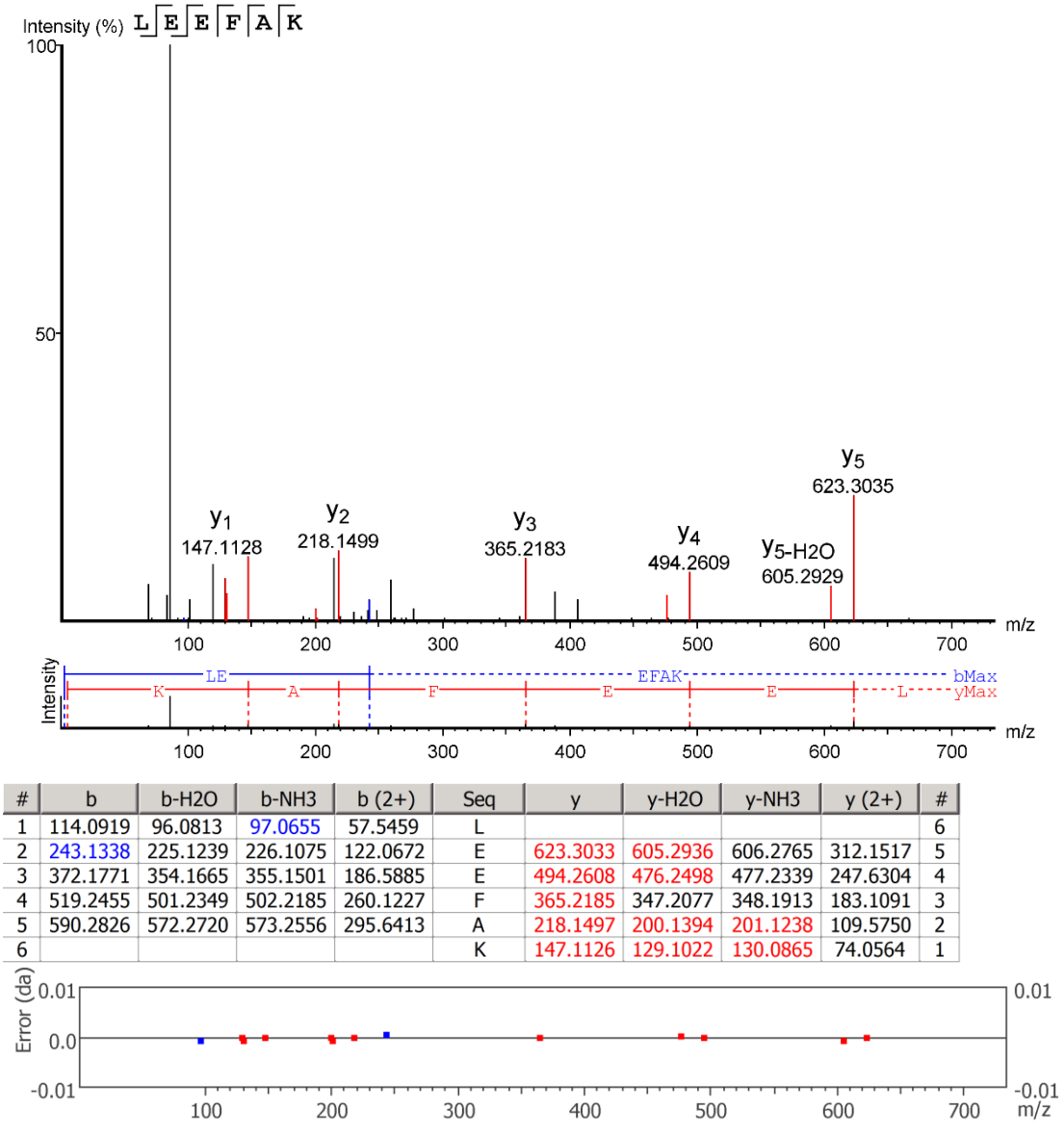


Figure 4.23 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. agrestis*** **tryptic peptide t8.**



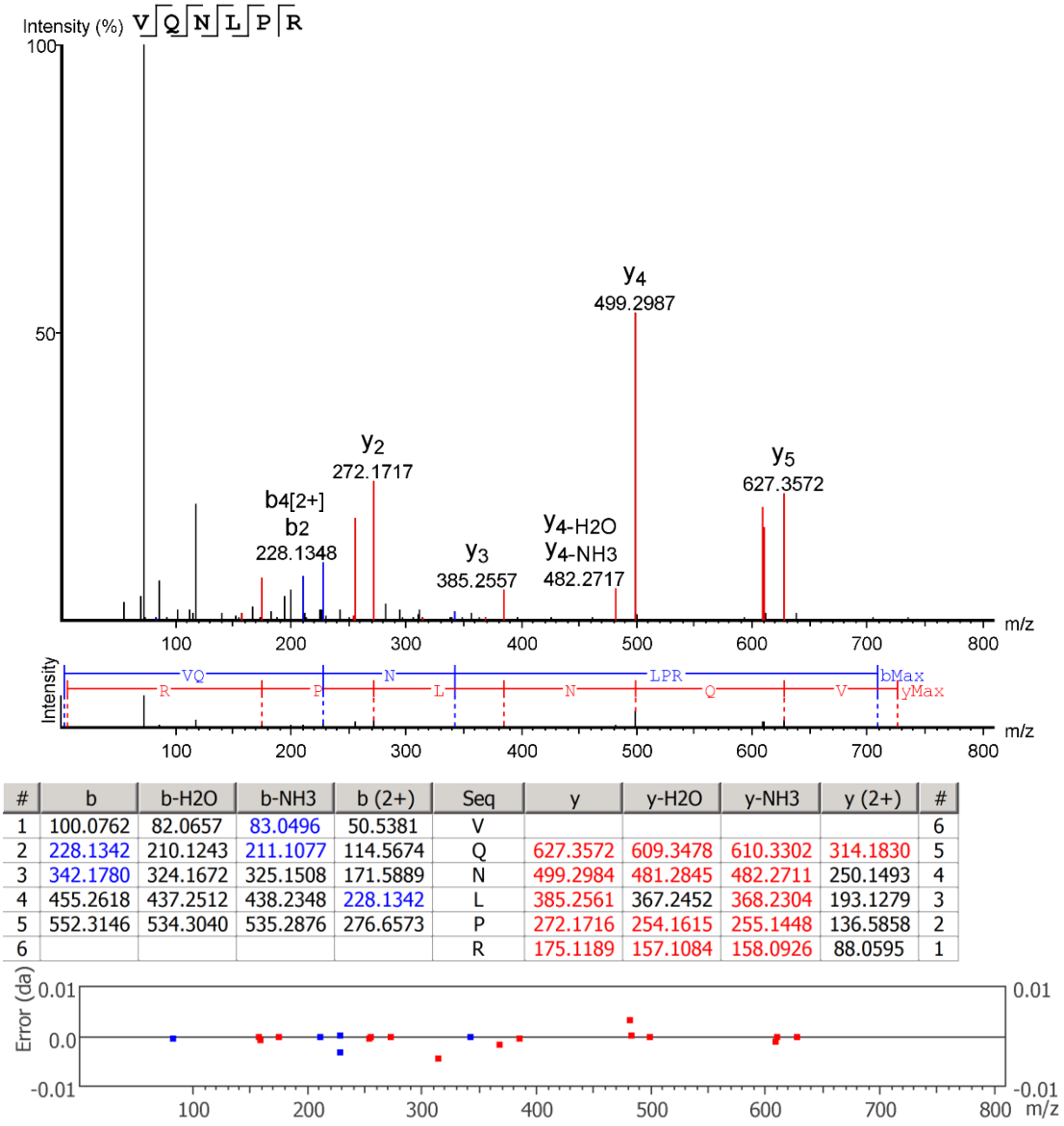
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					11
2	244.0927	226.0821	227.0661	122.5467	Q	1196.5900	1178.5729	1179.5636	598.7953	10
3	381.1515	363.1407	364.1247	191.0761	H	1068.5323	1050.5145	1051.5050	534.7660	9
4	494.2354	476.2253	477.2083	247.6181	L	931.4728	913.4629	914.4461	466.2412	8
5	595.2821	577.2726	578.2570	298.1420	T	818.3877	800.3772	801.3620	409.6945	7
6	692.3367	674.3262	675.3098	346.6684	P	717.3406	699.3296	700.3208	359.1707	6
7	821.3810	803.3702	804.3524	411.1917	E	620.2888	602.2770	603.2616	310.6443	5
8	949.4303	931.4274	932.4109	475.2227	Q	491.2458	473.2352	474.2165	246.1230	4
9	1036.4625	1018.4602	1019.4430	518.7350	S	363.1873	345.1763	346.1604	182.0937	3
10	1165.5077	1147.4973	1148.4856	583.2563	E	276.1555	258.1446	259.1284	138.5777	2
11					K	147.1127	129.1023	130.0865	74.0564	1

Figure 4.24 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t9.



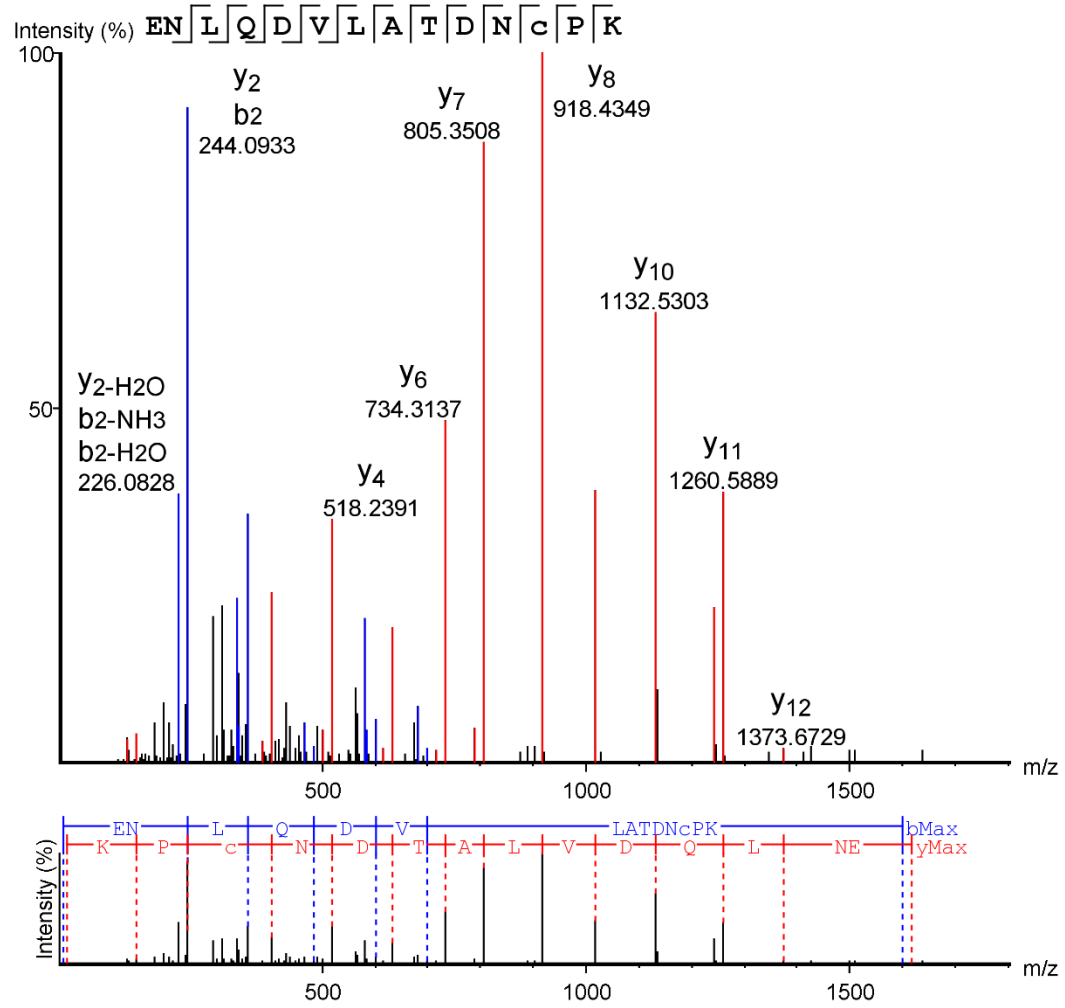
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0655	57.5459	L					6
2	243.1338	225.1239	226.1075	122.0672	E	623.3033	605.2936	606.2765	312.1517	5
3	372.1771	354.1665	355.1501	186.5885	E	494.2608	476.2498	477.2339	247.6304	4
4	519.2455	501.2349	502.2185	260.1227	F	365.2185	347.2077	348.1913	183.1091	3
5	590.2826	572.2720	573.2556	295.6413	A	218.1497	200.1394	201.1238	109.5750	2
6					K	147.1126	129.1022	130.0865	74.0564	1

Figure 4.25 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t10.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	100.0762	82.0657	83.0496	50.5381	V					6
2	228.1342	210.1243	211.1077	114.5674	Q	627.3572	609.3478	610.3302	314.1830	5
3	342.1780	324.1672	325.1508	171.5889	N	499.2984	481.2845	482.2711	250.1493	4
4	455.2618	437.2512	438.2348	228.1342	L	385.2561	367.2452	368.2304	193.1279	3
5	552.3146	534.3040	535.2876	276.6573	P	272.1716	254.1615	255.1448	136.5858	2
6					R	175.1189	157.1084	158.0926	88.0595	1

Figure 4.26 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* tryptic peptide t11.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	130.0504	112.0399	113.0234	65.5252	E					14
2	244.0924	226.0823	227.0666	122.5467	N	1487.7158	1469.7053	1470.6888	744.3579	13
3	357.1764	339.1665	340.1503	179.0887	L	1373.6766	1355.6624	1356.6459	687.3364	12
4	485.2338	467.2252	468.2090	243.1180	Q	1260.5884	1242.5784	1243.5625	630.7944	11
5	600.2626	582.2512	583.2404	300.6315	D	1132.5298	1114.5198	1115.5033	566.7651	10
6	699.3279	681.3212	682.3054	350.1656	V	1017.5037	999.4927	1000.4763	509.2516	9
7	812.4154	794.4048	795.3884	406.7077	L	918.4342	900.4244	901.4080	459.7175	8
8	883.4525	865.4420	866.4255	442.2263	A	805.3499	787.3391	788.3320	403.1754	7
9	984.5002	966.4896	967.4732	492.7501	T	734.3127	716.3029	717.2852	367.6569	6
10	1099.5271	1081.5166	1082.5001	550.2635	D	633.2657	615.2555	616.2382	317.1330	5
11	1213.5701	1195.5596	1196.5431	607.2850	N	518.2394	500.2286	501.2120	259.6196	4
12	1373.6007	1355.5902	1356.5737	687.3004	C(+57.02)	404.1964	386.1855	387.1699	202.5981	3
13	1470.6534	1452.6429	1453.6265	735.8267	P	244.1654	226.1548	227.1385	122.5828	2
14					K	147.1127	129.1024	130.0862	74.0564	1

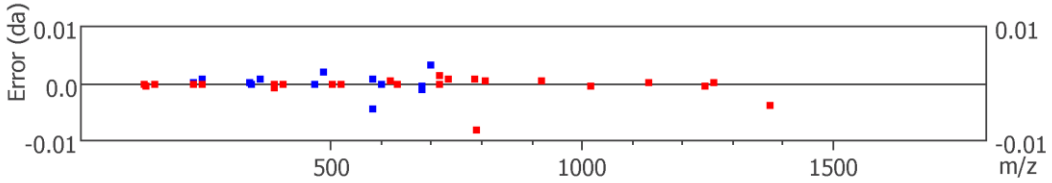
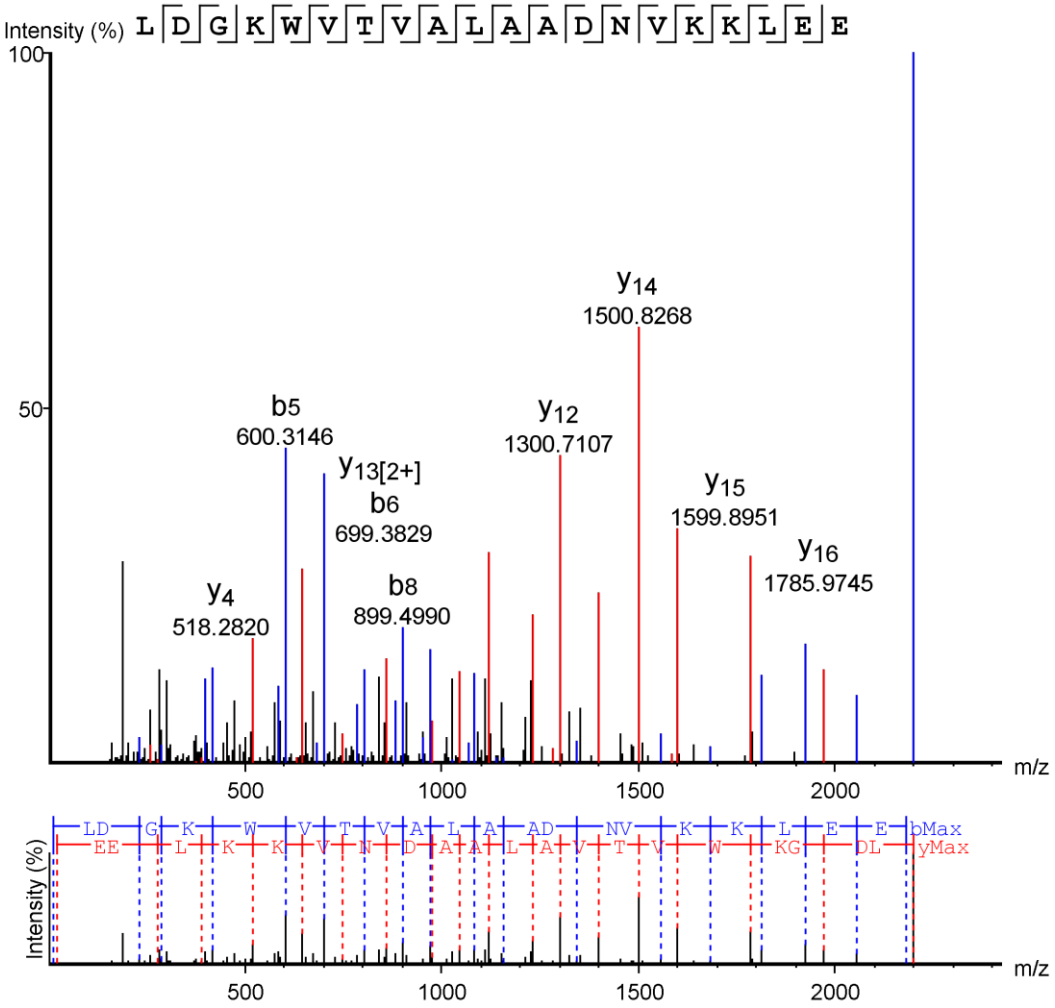


Figure 4.27 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* trytic peptide t12.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					20
2	229.1183	211.1083	212.0918	115.0594	D	2086.1179	2068.1074	2069.0908	1043.5613	19
3	286.1397	268.1297	269.1133	143.5701	G	1971.0875	1953.0803	1954.0638	986.0454	18
4	414.2342	396.2238	397.2083	207.6176	K	1914.0695	1896.0590	1897.0425	957.5347	17
5	600.3138	582.3040	583.2876	300.6573	W	1785.9712	1767.9640	1768.9475	893.4872	16
6	699.3813	681.3713	682.3560	350.1915	V	1599.8950	1581.8872	1582.8682	800.4476	15
7	800.4298	782.4178	783.4037	400.7153	T	1500.8256	1482.8163	1483.7998	750.9134	14
8	899.4983	881.4879	882.4720	450.2495	V	1399.7754	1381.7686	1382.7521	700.3846	13
9	970.5373	952.5226	953.5092	485.7681	A	1300.7100	1282.6986	1283.6837	650.8553	12
10	1083.6168	1065.6027	1066.5933	542.3101	L	1229.6738	1211.6631	1212.6466	615.3368	11
11	1154.6631	1136.6534	1137.6304	577.8287	A	1116.5890	1098.5790	1099.5625	558.7947	10
12	1225.6945	1207.6840	1208.6675	613.3472	A	1045.5526	1027.5419	1028.5254	523.2762	9
13	1340.7202	1322.7109	1323.6945	670.8607	D	974.5189	956.5047	957.4883	487.7576	8
14	1454.7643	1436.7538	1437.7373	727.8821	N	859.4866	841.4777	842.4613	430.2441	7
15	1553.8335	1535.8223	1536.8058	777.4164	V	745.4441	727.4348	728.4184	373.2227	6
16	1681.9233	1663.9172	1664.9008	841.4639	K	646.3769	628.3627	629.3500	323.6885	5
17	1810.0189	1792.0121	1792.9956	905.5113	K	518.2820	500.2715	501.2550	259.6410	4
18	1923.1083	1905.0962	1906.0797	962.0533	L	390.1866	372.1765	373.1546	195.5935	3
19	2052.1511	2034.1387	2035.1222	1026.5795	E	277.1028	259.0923	260.0760	139.0515	2
20					E	148.0604	130.0499	131.0334	74.5302	1

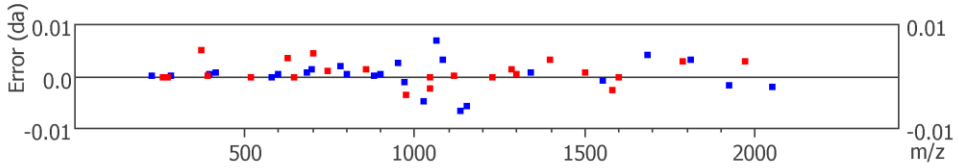
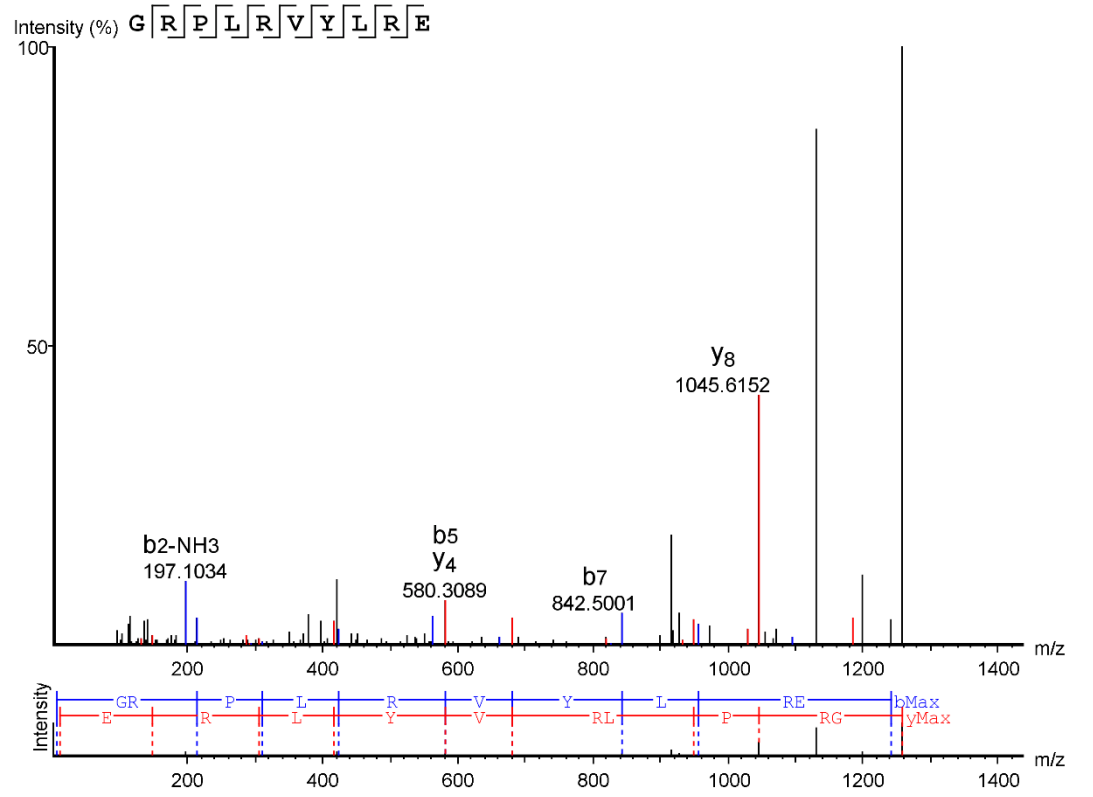


Figure 4.28 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g1.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					10
2	214.1298	196.1198	197.1034	107.5652	R	1201.7163	1183.7058	1184.6896	601.3582	9
3	311.1826	293.1726	294.1548	156.0916	P	1045.6152	1027.6047	1028.5911	523.3076	8
4	424.2661	406.2567	407.2402	212.6336	L	948.5630	930.5519	931.5313	474.7812	7
5	580.3667	562.3577	563.3409	290.6841	R	835.4784	817.4678	818.4493	418.2392	6
6	679.4327	661.4261	662.4081	340.2184	V	679.3765	661.3668	662.3503	340.1887	5
7	842.4982	824.4895	825.4771	421.7500	Y	580.3080	562.2983	563.2819	290.6588	4
8	955.5810	937.5735	938.5571	478.2921	L	417.2460	399.2350	400.2186	209.1228	3
9	1111.6852	1093.6747	1094.6580	556.3426	R	304.1610	286.1509	287.1348	152.5807	2
10					E	148.0603	130.0501	131.0334	74.5302	1

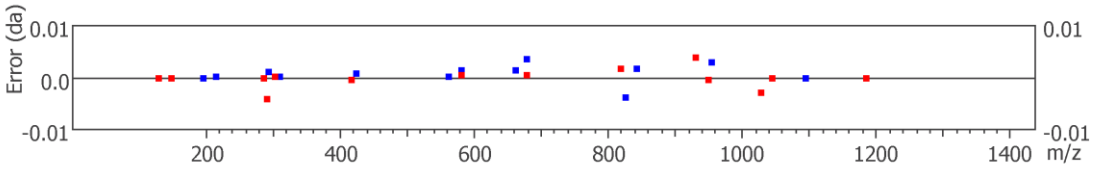
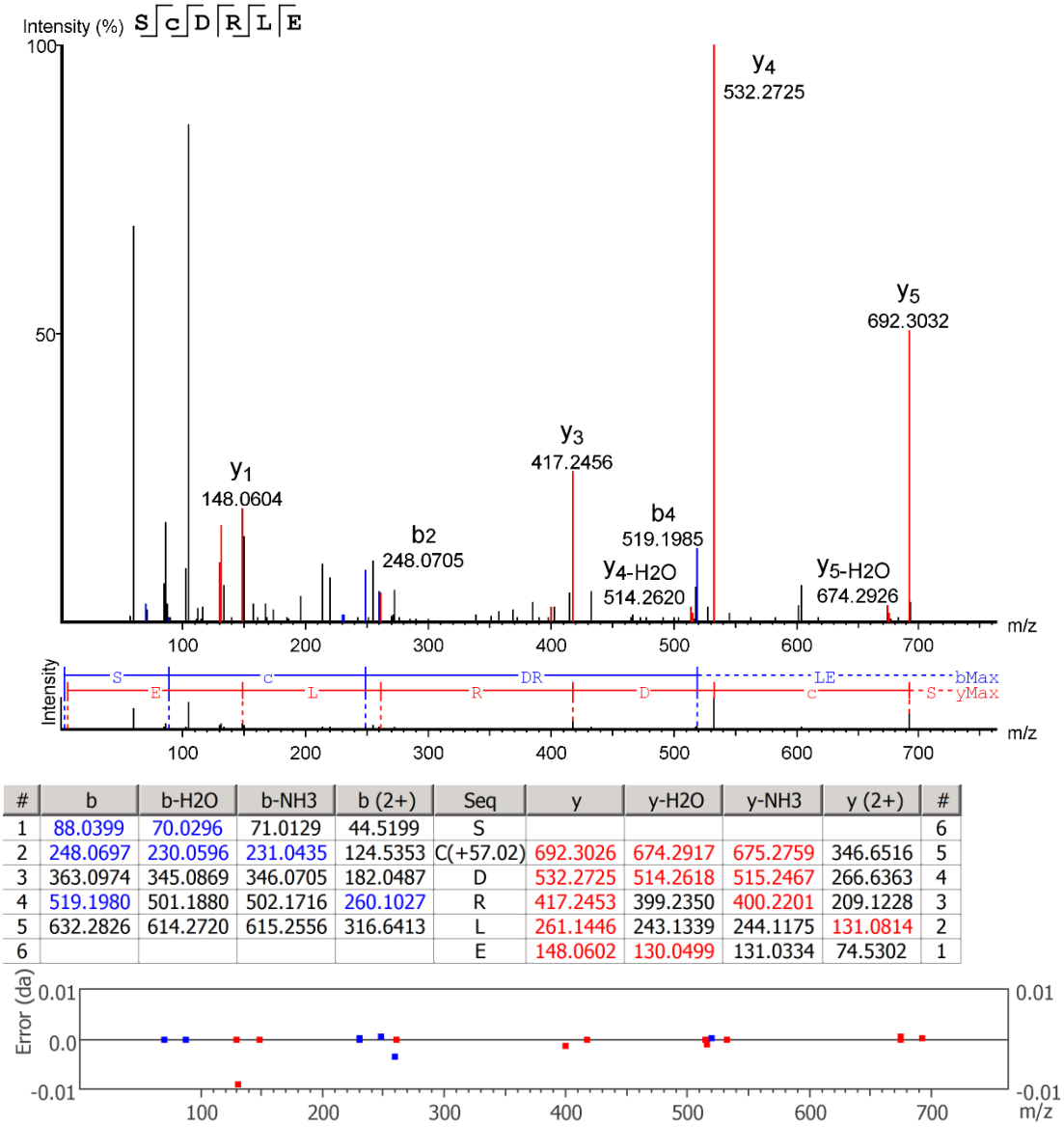
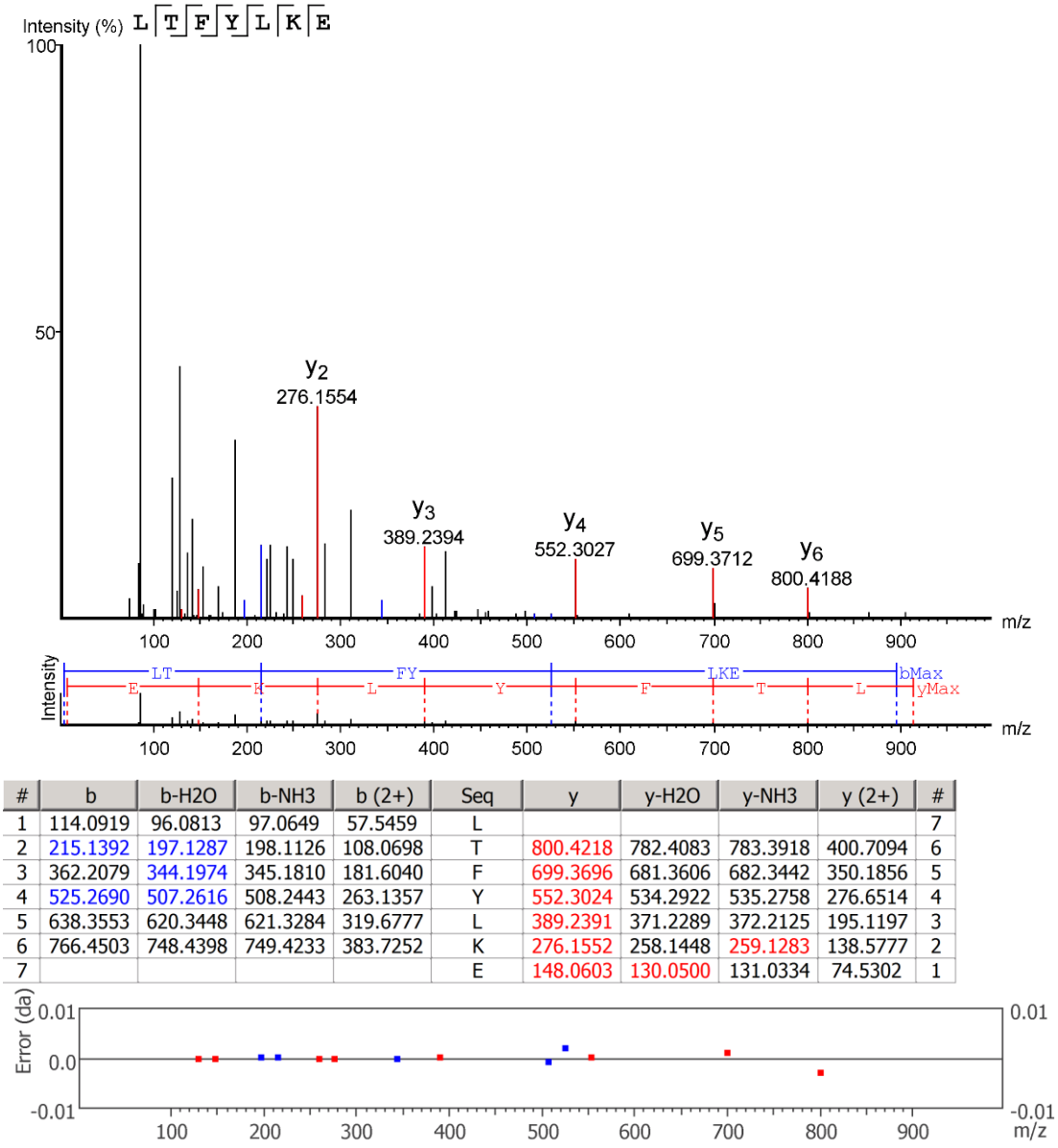


Figure 4.29 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g2.



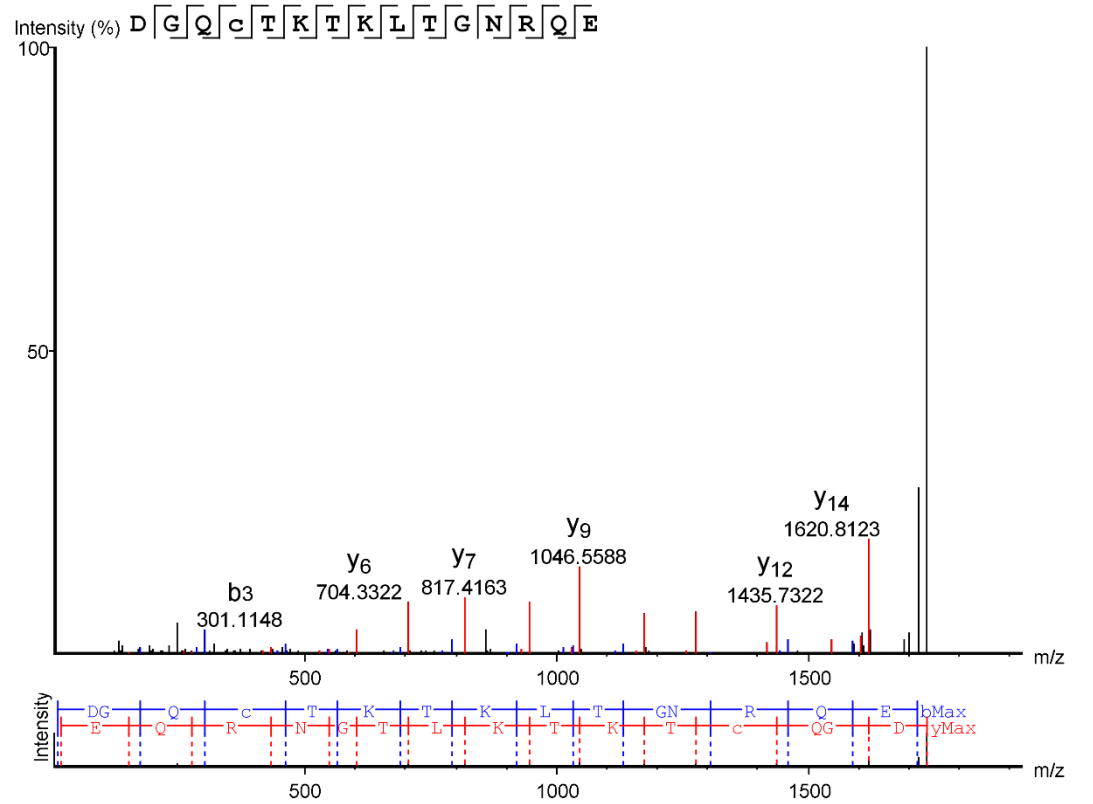
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	88.0399	70.0296	71.0129	44.5199	S					6
2	248.0697	230.0596	231.0435	124.5353	C(+57.02)	692.3026	674.2917	675.2759	346.6516	5
3	363.0974	345.0869	346.0705	182.0487	D	532.2725	514.2618	515.2467	266.6363	4
4	519.1980	501.1880	502.1716	260.1027	R	417.2453	399.2350	400.2201	209.1228	3
5	632.2826	614.2720	615.2556	316.6413	L	261.1446	243.1339	244.1175	131.0814	2
6					E	148.0602	130.0499	131.0334	74.5302	1

Figure 4.30 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g3.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					7
2	215.1392	197.1287	198.1126	108.0698	T	800.4218	782.4083	783.3918	400.7094	6
3	362.2079	344.1974	345.1810	181.6040	F	699.3696	681.3606	682.3442	350.1856	5
4	525.2690	507.2616	508.2443	263.1357	Y	552.3024	534.2922	535.2758	276.6514	4
5	638.3553	620.3448	621.3284	319.6777	L	389.2391	371.2289	372.2125	195.1197	3
6	766.4503	748.4398	749.4233	383.7252	K	276.1552	258.1448	259.1283	138.5777	2
7					E	148.0603	130.0500	131.0334	74.5302	1

Figure 4.31 ***De novo*** **sequence analysis of the processed MS/MS spectra of** ***M. agrestis*** **glu-C peptide g4.**



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					15
2	173.0557	155.0457	156.0292	87.0281	G	1620.8085	1602.8018	1603.7920	810.9061	14
3	301.1140	283.1036	284.0873	151.0574	Q	1563.7908	1545.7803	1546.7650	782.3954	13
4	461.1448	443.1334	444.1161	231.0727	C(+57.02)	1435.7334	1417.7217	1418.7092	718.3661	12
5	562.1915	544.1809	545.1661	281.5966	T	1275.7006	1257.6910	1258.6776	638.3508	11
6	690.2883	672.2769	673.2611	345.6440	K	1174.6521	1156.6426	1157.6301	587.8269	10
7	791.3352	773.3251	774.3088	396.1679	T	1046.5571	1028.5444	1029.5347	523.7794	9
8	919.4305	901.4172	902.3997	460.2154	K	945.5107	927.5006	928.4850	473.2556	8
9	1032.5103	1014.5040	1015.4878	516.7574	L	817.4155	799.4057	800.3893	409.2081	7
10	1133.5588	1115.5587	1116.5355	567.2812	T	704.3318	686.3216	687.3059	352.6661	6
11	1190.5840	1172.5735	1173.5570	595.7920	G	603.2844	585.2739	586.2573	302.1422	5
12	1304.6198	1286.6163	1287.5999	652.8134	N	546.2626	528.2524	529.2366	273.6315	4
13	1460.7214	1442.7175	1443.7030	730.8640	R	432.2199	414.2095	415.1940	216.6100	3
14	1588.7821	1570.7760	1571.7595	794.8932	Q	276.1190	258.1080	259.0923	138.5595	2
15					E	148.0602	130.0499	131.0334	74.5302	1

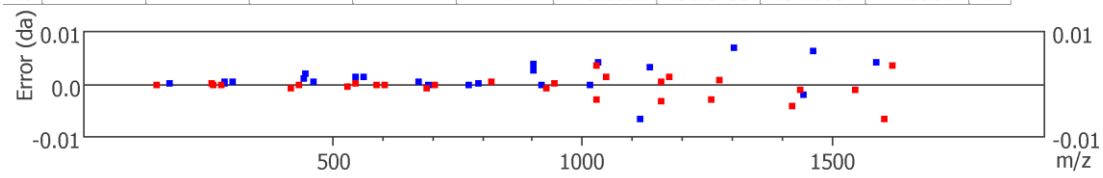
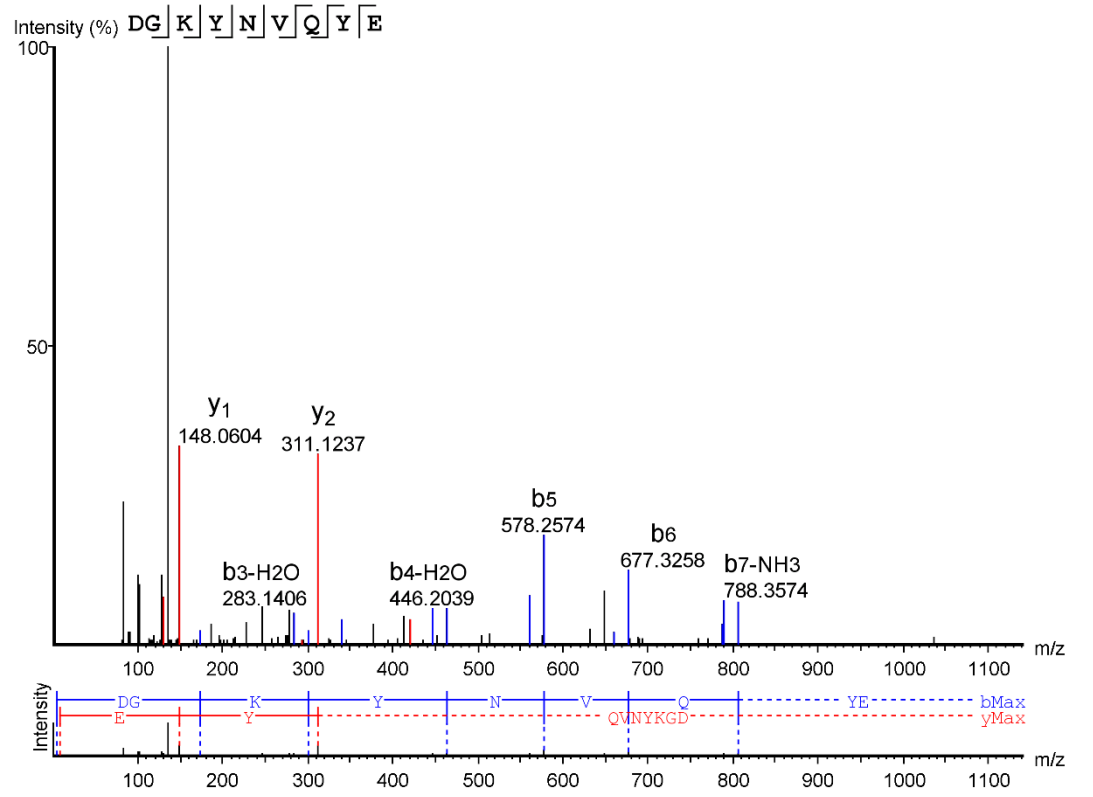


Figure 4.32 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g5.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					9
2	173.0553	155.0457	156.0292	87.0281	G	1000.4734	982.4628	983.4464	500.7367	8
3	301.1506	283.1399	284.1242	151.0756	K	943.4519	925.4413	926.4249	472.2260	7
4	464.2133	446.2032	447.1875	232.6072	Y	815.3570	797.3464	798.3300	408.1785	6
5	578.2567	560.2434	561.2294	289.6287	N	652.2936	634.2831	635.2667	326.6468	5
6	677.3246	659.3121	660.2976	339.1658	V	538.2507	520.2402	521.2238	269.6254	4
7	805.3817	787.3741	788.3580	403.1922	Q	439.1823	421.1715	422.1553	220.0912	3
8	968.4478	950.4372	951.4208	484.7239	Y	311.1233	293.1122	294.0967	156.0619	2
9					E	148.0604	130.0502	131.0334	74.5302	1

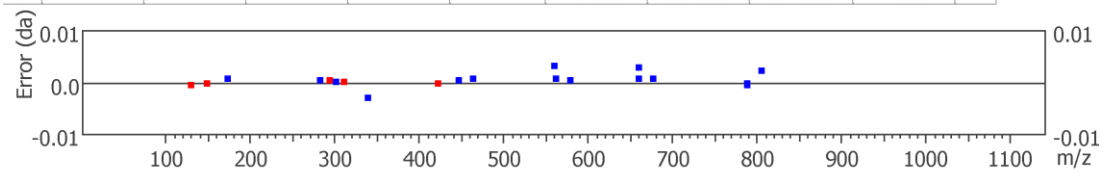
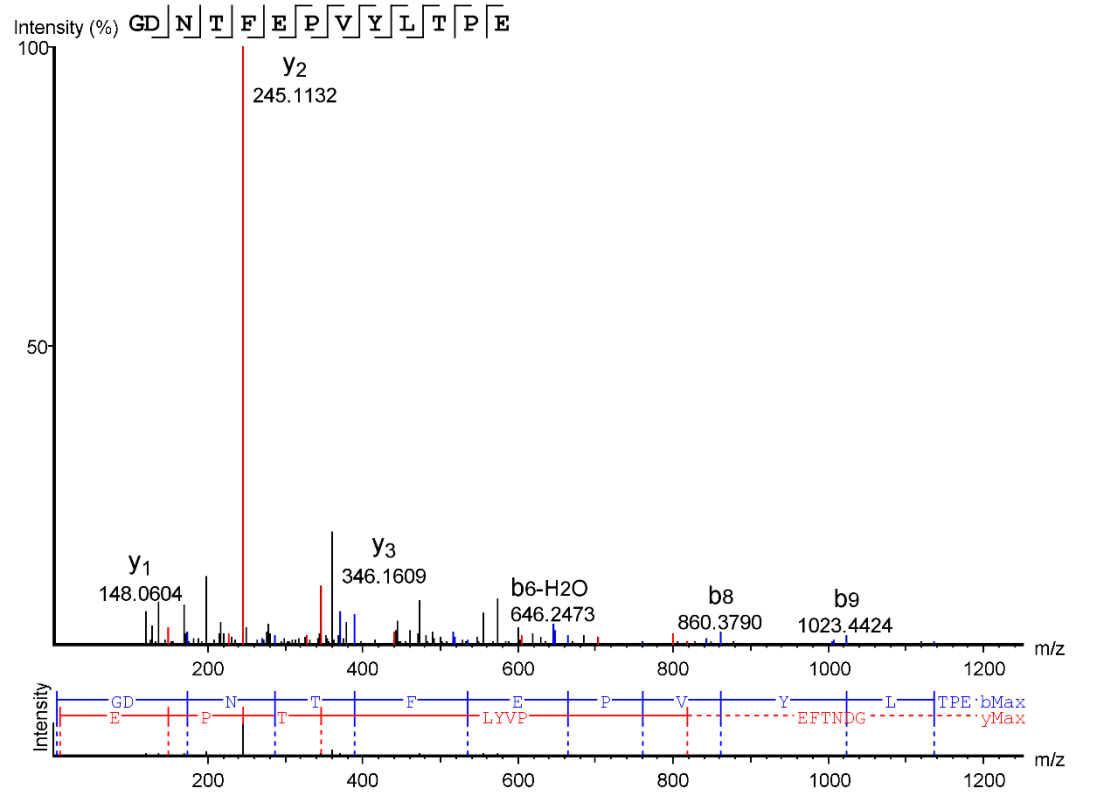


Figure 4.33 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g6.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					13
2	173.0558	155.0457	156.0292	87.0281	D	1424.6580	1406.6475	1407.6310	712.8290	12
3	287.0988	269.0875	270.0722	144.0496	N	1309.6310	1291.6205	1292.6040	655.3155	11
4	388.1459	370.1354	371.1180	194.5734	T	1195.5881	1177.5776	1178.5612	598.2941	10
5	535.2151	517.2041	518.1903	268.1076	F	1094.5404	1076.5299	1077.5134	547.7702	9
6	664.2563	646.2449	647.2277	332.6289	E	947.4720	929.4614	930.4450	474.2360	8
7	761.3149	743.3000	744.2836	381.1553	P	818.4282	800.4190	801.4024	409.7147	7
8	860.3727	842.3694	843.3572	430.6895	V	721.3766	703.3662	704.3497	361.1883	6
9	1023.4366	1005.4254	1006.4169	512.2212	Y	622.3082	604.2964	605.2812	311.6541	5
10	1136.5205	1118.5159	1119.4994	568.7632	L	459.2449	441.2339	442.2179	230.1225	4
11	1237.5741	1219.5636	1220.5471	619.2870	T	346.1606	328.1501	329.1339	173.5804	3
12	1334.6268	1316.6163	1317.5999	667.8134	P	245.1131	227.1024	228.0862	123.0566	2
13					E	148.0603	130.0499	131.0334	74.5302	1

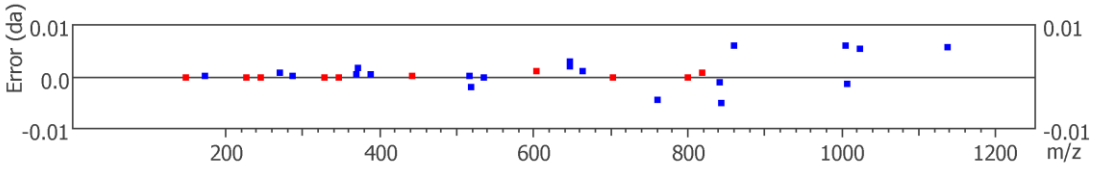
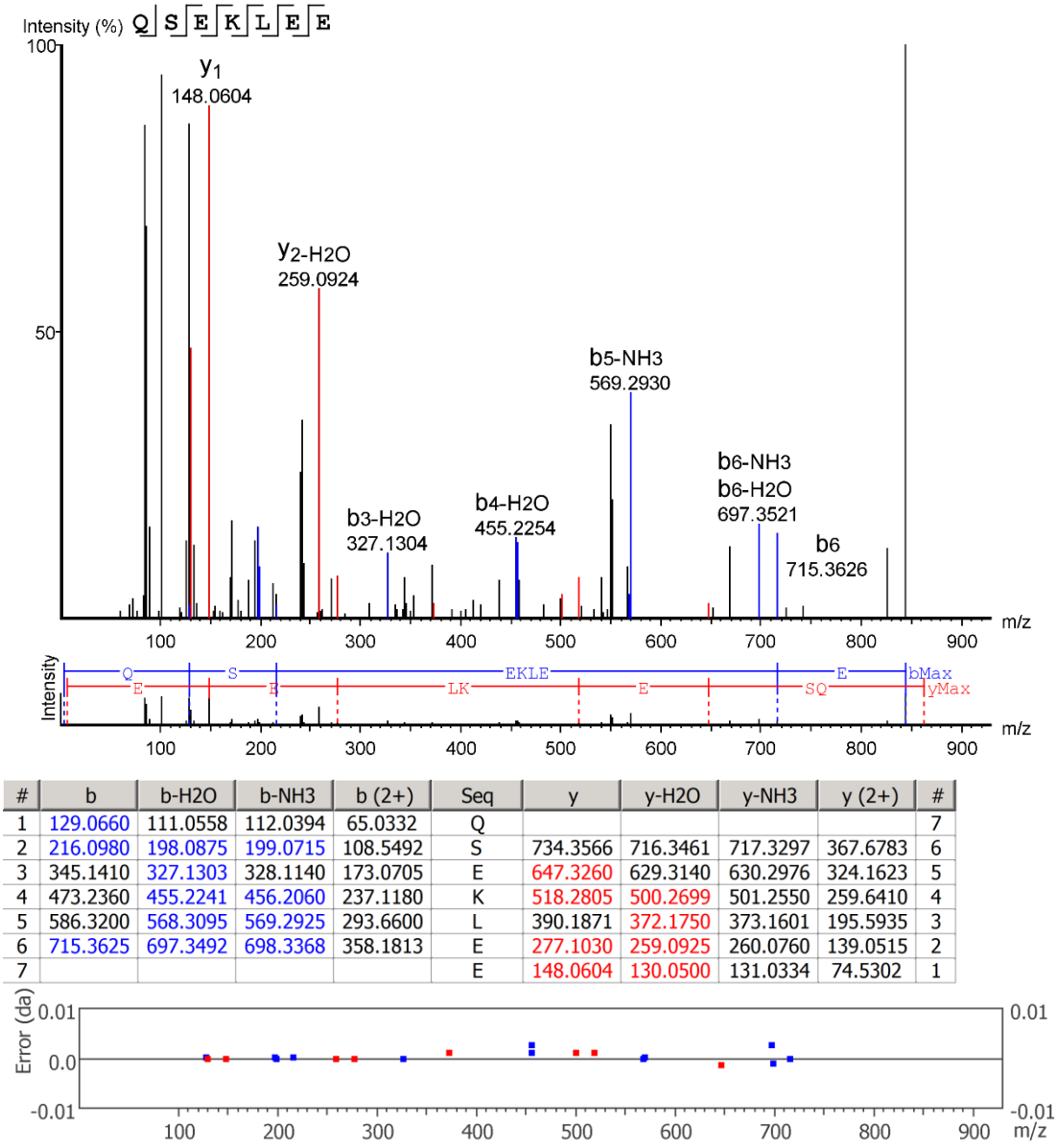
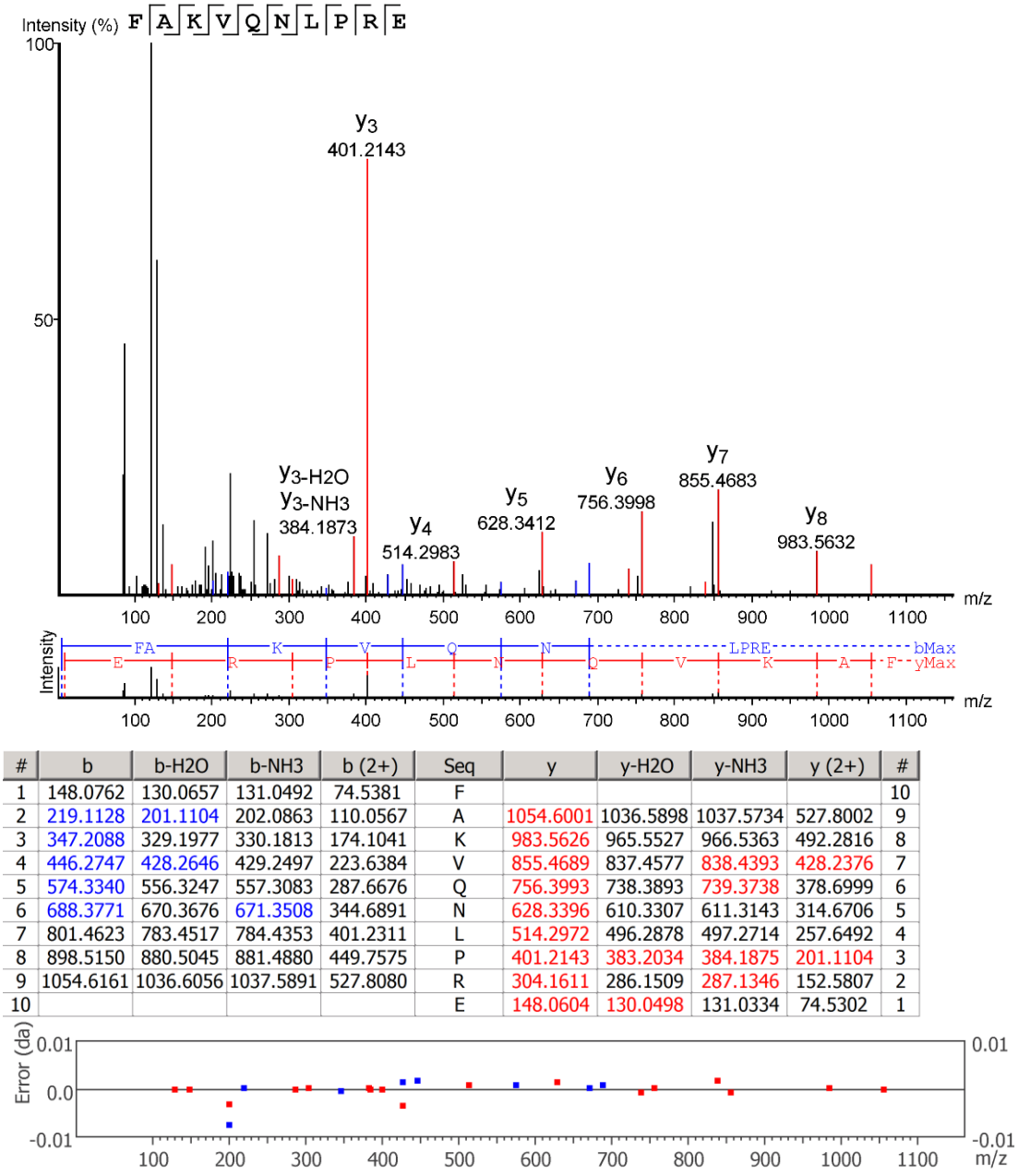


Figure 4.34 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g7.



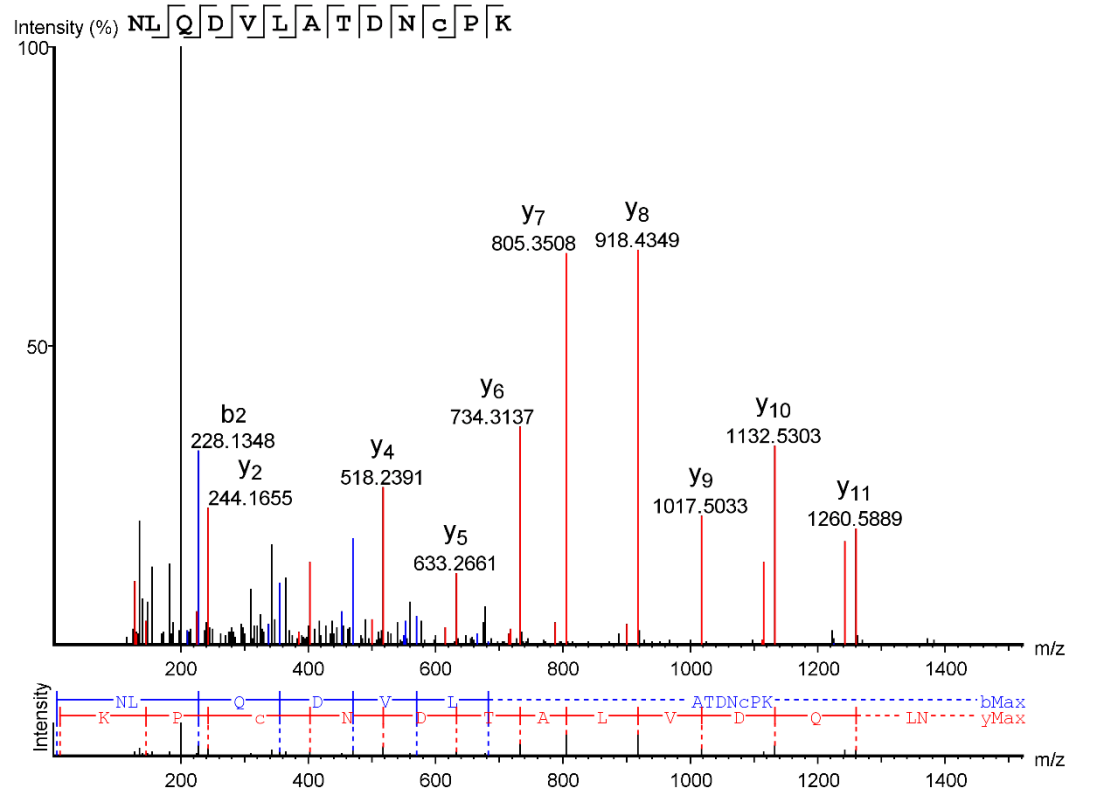
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	129.0660	111.0558	112.0394	65.0332	Q					7
2	216.0980	198.0875	199.0715	108.5492	S	734.3566	716.3461	717.3297	367.6783	6
3	345.1410	327.1303	328.1140	173.0705	E	647.3260	629.3140	630.2976	324.1623	5
4	473.2360	455.2241	456.2060	237.1180	K	518.2805	500.2699	501.2550	259.6410	4
5	586.3200	568.3095	569.2925	293.6600	L	390.1871	372.1750	373.1601	195.5935	3
6	715.3625	697.3492	698.3368	358.1813	E	277.1030	259.0925	260.0760	139.0515	2
7					E	148.0604	130.0500	131.0334	74.5302	1

Figure 4.35 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g8.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	148.0762	130.0657	131.0492	74.5381	F					10
2	219.1128	201.1104	202.0863	110.0567	A	1054.6001	1036.5898	1037.5734	527.8002	9
3	347.2088	329.1977	330.1813	174.1041	K	983.5626	965.5527	966.5363	492.2816	8
4	446.2747	428.2646	429.2497	223.6384	V	855.4689	837.4577	838.4393	428.2376	7
5	574.3340	556.3247	557.3083	287.6676	Q	756.3993	738.3893	739.3738	378.6999	6
6	688.3771	670.3676	671.3508	344.6891	N	628.3396	610.3307	611.3143	314.6706	5
7	801.4623	783.4517	784.4353	401.2311	L	514.2972	496.2878	497.2714	257.6492	4
8	898.5150	880.5045	881.4880	449.7575	P	401.2143	383.2034	384.1875	201.1104	3
9	1054.6161	1036.6056	1037.5891	527.8080	R	304.1611	286.1509	287.1346	152.5807	2
10					E	148.0604	130.0498	131.0334	74.5302	1

Figure 4.36 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g9.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	115.0508	97.0402	98.0238	58.0254	N					13
2	228.1344	210.1243	211.1077	114.5674	L	1373.6729	1355.6624	1356.6459	687.3364	12
3	356.1927	338.1822	339.1669	178.5967	Q	1260.5876	1242.5768	1243.5669	630.7944	11
4	471.2195	453.2089	454.1911	236.1102	D	1132.5299	1114.5166	1115.5013	566.7651	10
5	570.2888	552.2770	553.2606	285.6444	V	1017.5027	999.4927	1000.4763	509.2516	9
6	683.3723	665.3585	666.3510	342.1864	L	918.4338	900.4254	901.4103	459.7175	8
7	754.4099	736.3994	737.3829	377.7050	A	805.3492	787.3392	788.3239	403.1754	7
8	855.4576	837.4470	838.4306	428.2288	T	734.3135	716.3052	717.2863	367.6569	6
9	970.4845	952.4739	953.4575	485.7422	D	633.2654	615.2560	616.2394	317.1330	5
10	1084.5275	1066.5170	1067.5005	542.7637	N	518.2391	500.2286	501.2120	259.6196	4
11	1244.5581	1226.5442	1227.5311	622.7791	C(+57.02)	404.1957	386.1844	387.1682	202.5981	3
12	1341.6108	1323.6003	1324.5839	671.3054	P	244.1654	226.1550	227.1385	122.5828	2
13					K	147.1126	129.1025	130.0863	74.0564	1

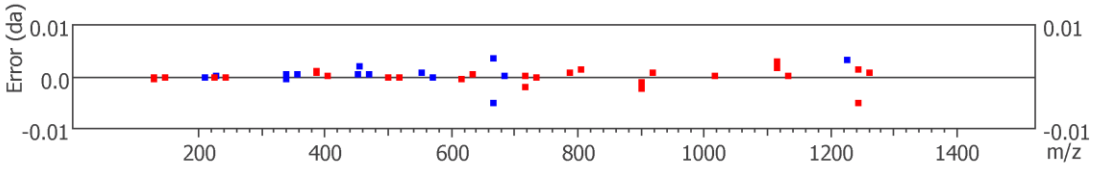
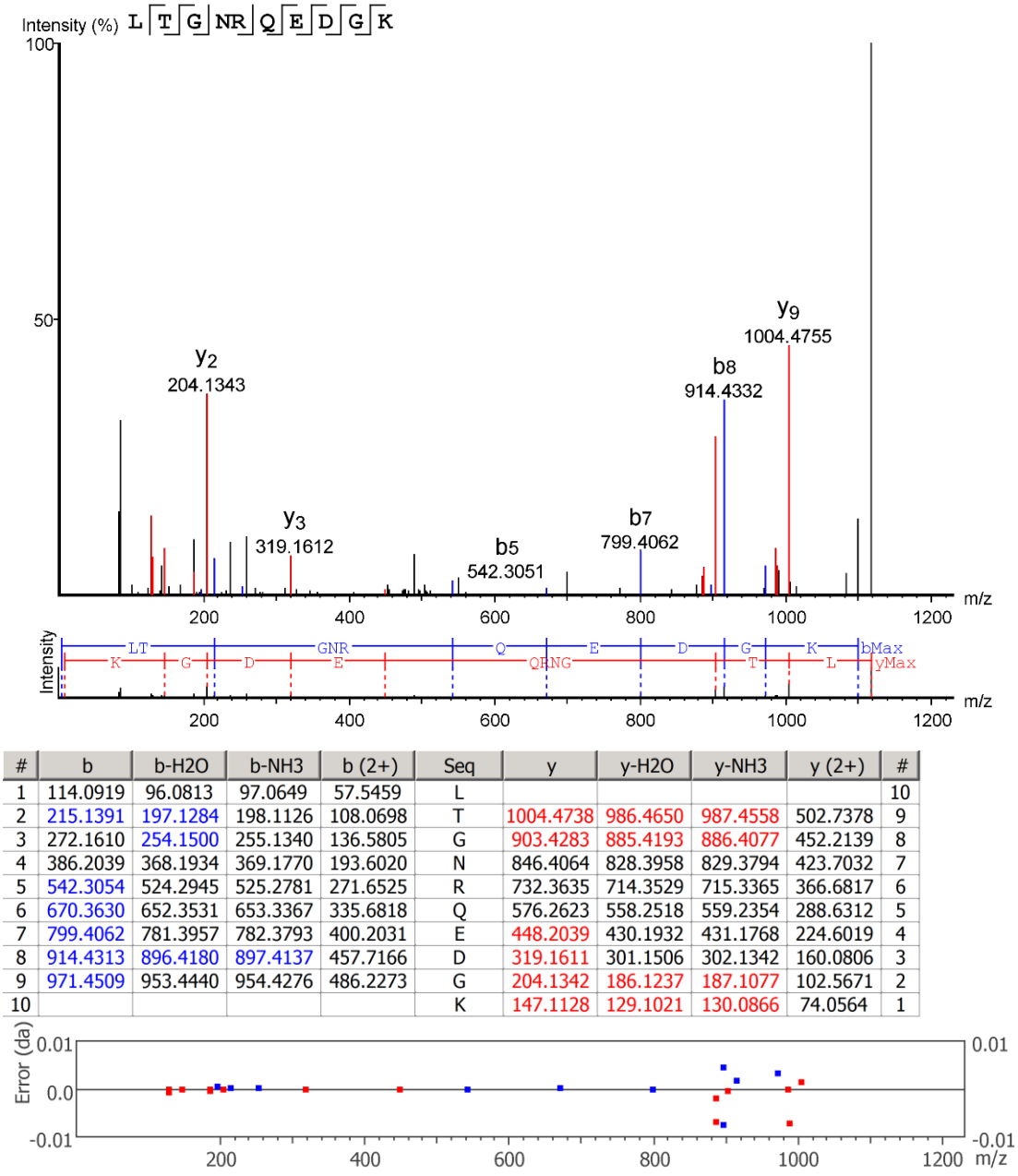
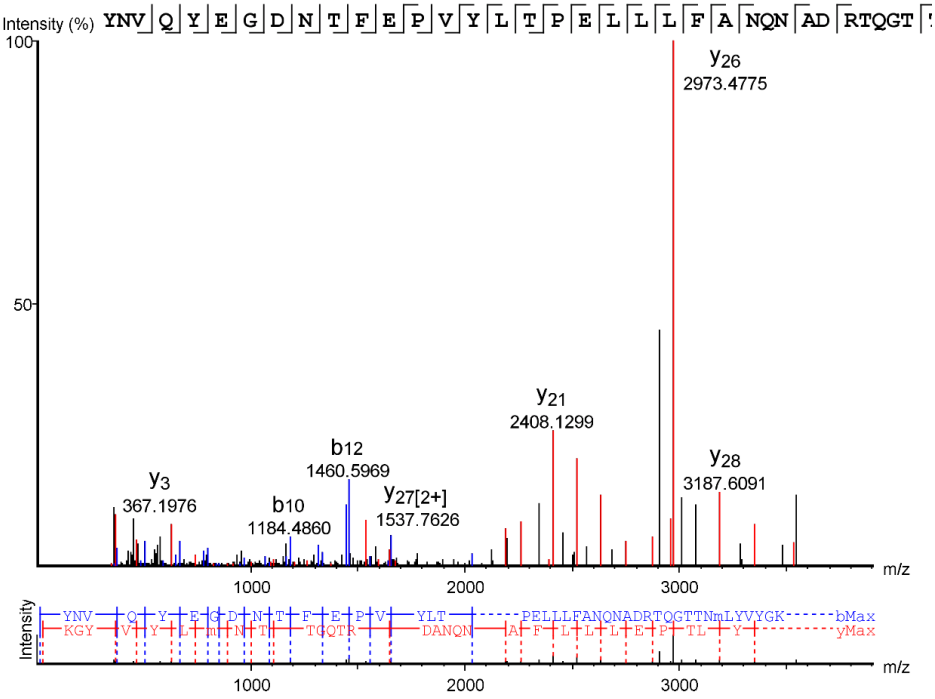


Figure 4.37 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* glu-C peptide g10.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					10
2	215.1391	197.1284	198.1126	108.0698	T	1004.4738	986.4650	987.4558	502.7378	9
3	272.1610	254.1500	255.1340	136.5805	G	903.4283	885.4193	886.4077	452.2139	8
4	386.2039	368.1934	369.1770	193.6020	N	846.4064	828.3958	829.3794	423.7032	7
5	542.3054	524.2945	525.2781	271.6525	R	732.3635	714.3529	715.3365	366.6817	6
6	670.3630	652.3531	653.3367	335.6818	Q	576.2623	558.2518	559.2354	288.6312	5
7	799.4062	781.3957	782.3793	400.2031	E	448.2039	430.1932	431.1768	224.6019	4
8	914.4313	896.4180	897.4137	457.7166	D	319.1611	301.1506	302.1342	160.0806	3
9	971.4509	953.4440	954.4276	486.2273	G	204.1342	186.1237	187.1077	102.5671	2
10					K	147.1128	129.1021	130.0866	74.0564	1

Figure 4.38 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* lys-C peptide l1.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	164.0712	146.0606	147.0441	82.5356	Y					43
2	278.1141	260.1035	261.0871	139.5570	N	4843.3198	4825.3091	4826.2925	2422.1599	42
3	377.1819	359.1719	360.1553	189.0912	V	4729.2769	4711.2661	4712.2495	2365.1384	41
4	505.2404	487.2305	488.2147	253.1205	Q	4630.2085	4612.1978	4613.1812	2315.6042	40
5	668.3022	650.2935	651.2747	334.6522	Y	4502.1499	4484.1392	4485.1226	2251.5750	39
6	797.3466	779.3364	780.3210	399.1735	E	4339.0864	4321.0757	4322.0591	2170.0432	38
7	854.3680	836.3578	837.3405	427.6842	G	4210.0439	4192.0332	4193.0166	2105.5220	37
8	969.3926	951.3768	952.3684	485.1977	D	4153.0225	4135.0117	4135.9951	2077.0112	36
9	1083.4382	1065.4279	1066.4054	542.2192	N	4037.9954	4019.9849	4020.9683	2019.4977	35
10	1184.4871	1166.4755	1167.4629	592.7441	T	3923.9524	3905.9419	3906.9253	1962.4762	34
11	1331.5513	1313.5376	1314.5322	666.2772	F	3822.9045	3804.8940	3805.8774	1911.9523	33
12	1460.5999	1442.5809	1443.5751	730.7985	E	3675.8362	3657.8257	3658.8091	1838.4181	32
13	1557.6509	1539.6393	1540.6228	779.3249	P	3546.7937	3528.7832	3529.7761	1773.8969	31
14	1656.7103	1638.7102	1639.6881	828.8591	V	3449.7410	3431.7305	3432.7139	1725.3705	30
15	1819.7815	1801.7710	1802.7545	910.3907	Y	3350.6667	3332.6621	3333.6455	1675.8271	29
16	1932.8655	1914.8550	1915.8385	966.9327	L	3187.6155	3169.5986	3170.5820	1594.3116	28
17	2033.9043	2015.9027	2016.8862	1017.4566	T	3074.5251	3056.5146	3057.4980	1537.7633	27
18	2130.9661	2112.9556	2113.9390	1065.9830	P	2973.4744	2955.4670	2956.4551	1487.2388	26
19	2260.0085	2241.9980	2242.9814	1130.5043	E	2876.4307	2858.4141	2859.3975	1438.7123	25
20	2373.0925	2355.0820	2356.0654	1187.0463	L	2747.3799	2729.3716	2730.3550	1374.1987	24
21	2486.1768	2468.1663	2469.1497	1243.5884	L	2634.2998	2616.2876	2617.2710	1317.6490	23
22	2599.2607	2581.2502	2582.2336	1300.1304	L	2521.2129	2503.2034	2504.1868	1261.1104	22
23	2746.3291	2728.3186	2729.3020	1373.6646	F	2408.1304	2390.1194	2391.0996	1204.5691	21
24	2817.3662	2799.3557	2800.3391	1409.1831	A	2261.0576	2243.0510	2244.0344	1131.0308	20
25	2931.4092	2913.3987	2914.3821	1466.2046	N	2190.0181	2172.0139	2172.9973	1095.5182	19
26	3059.4678	3041.4573	3042.4407	1530.2339	Q	2075.9814	2057.9709	2058.9543	1038.4907	18
27	3173.5107	3155.5002	3156.4836	1587.2554	N	1947.9229	1929.9124	1930.8959	974.4614	17
28	3244.5479	3226.5374	3227.5208	1622.7739	A	1833.8800	1815.8695	1816.8530	917.4355	16
29	3359.5747	3341.5642	3342.5476	1680.2874	D	1762.8429	1744.8324	1745.8159	881.9214	15
30	3515.6758	3497.6653	3498.6487	1758.3379	R	1647.8138	1629.8054	1630.7889	824.4124	14
31	3616.7236	3598.7131	3599.6965	1808.8618	T	1491.7148	1473.7043	1474.6879	746.3574	13
32	3744.7822	3726.7717	3727.7551	1872.8911	Q	1390.6671	1372.6566	1373.6401	695.8336	12
33	3801.8035	3783.7930	3784.7764	1901.4017	G	1262.6085	1244.5980	1245.5815	631.8043	11
34	3902.8513	3884.8408	3885.8242	1951.9257	T	1205.5870	1187.5765	1188.5601	603.2935	10
35	4003.8989	3985.8884	3986.8718	2002.4495	T	1104.5380	1086.5289	1087.5125	552.7697	9
36	4117.9419	4099.9312	4100.9150	2059.4709	N	1003.4862	985.4811	986.4647	502.2458	8
37	4264.9771	4246.9663	4247.9502	2132.9885	M(+15....	889.4478	871.4382	872.4218	445.2244	7
38	4378.0615	4360.0508	4361.0347	2189.5308	L	742.4119	724.4028	725.3864	371.7067	6
39	4541.1245	4523.1138	4524.0977	2271.0623	Y	629.3286	611.3187	612.3023	315.1646	5
40	4640.1929	4622.1821	4623.1660	2320.5964	V	466.2658	448.2554	449.2390	233.6330	4
41	4803.2563	4785.2456	4786.2295	2402.1282	Y	367.1976	349.1870	350.1706	184.0988	3
42	4860.2778	4842.2671	4843.2510	2430.6389	G	204.1343	186.1237	187.1073	102.5671	2
43					K	147.1128	129.1022	130.0858	74.0564	1

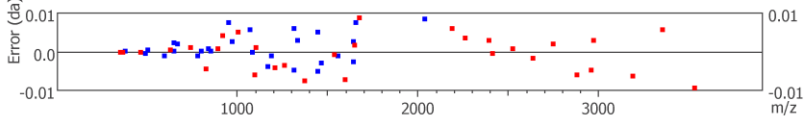
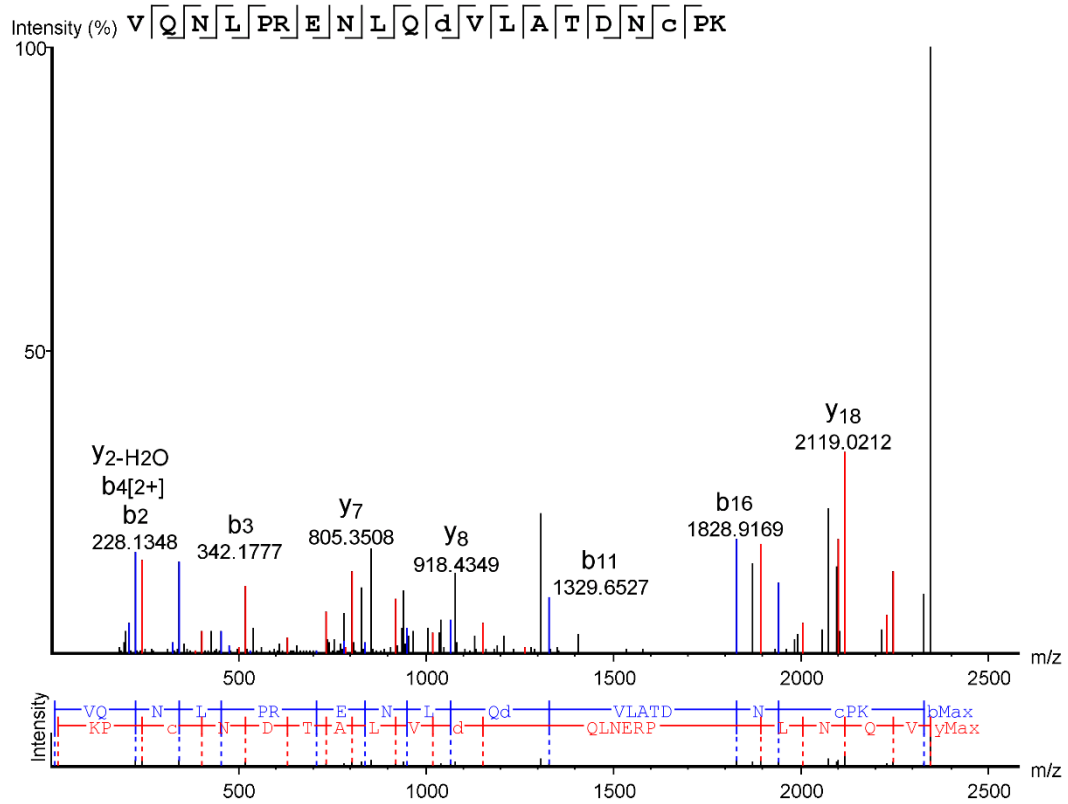


Figure 4.39 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* lys-C peptide l2.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	100.0762	82.0657	83.0492	50.5381	V					20
2	228.1344	210.1243	211.1078	114.5674	Q	2247.0825	2229.0693	2230.0591	1124.0399	19
3	342.1774	324.1672	325.1508	171.5889	N	2119.0208	2101.0107	2101.9873	1060.0106	18
4	455.2607	437.2512	438.2348	228.1344	L	2004.9750	1986.9678	1987.9513	1002.9891	17
5	552.3146	534.3040	535.2876	276.6573	P	1891.8938	1873.8838	1874.8673	946.4471	16
6	708.4161	690.4051	691.3887	354.7078	R	1794.8414	1776.8309	1777.8145	897.9207	15
7	837.4611	819.4477	820.4313	419.2291	E	1638.7404	1620.7299	1621.7134	819.8702	14
8	951.4932	933.4906	934.4742	476.2536	N	1509.6978	1491.6873	1492.6708	755.3489	13
9	1064.5848	1046.5747	1047.5582	532.7969	L	1395.6548	1377.6443	1378.6278	698.3274	12
10	1192.6438	1174.6333	1175.6168	596.8219	Q	1282.5708	1264.5603	1265.5400	641.7854	11
11	1329.6514	1311.6422	1312.6257	665.3264	D(+21.98)	1154.5128	1136.5017	1137.4852	577.7561	10
12	1428.7211	1410.7106	1411.6941	714.8605	V	1017.5009	999.4927	1000.4763	509.2516	9
13	1541.8052	1523.7947	1524.7782	771.4026	L	918.4340	900.4244	901.4080	459.7175	8
14	1612.8423	1594.8318	1595.8153	806.9211	A	805.3505	787.3341	788.3239	403.1754	7
15	1713.8900	1695.8795	1696.8630	857.4450	T	734.3129	716.3032	717.2867	367.6569	6
16	1828.9137	1810.9064	1811.8899	914.9584	D	633.2656	615.2555	616.2391	317.1330	5
17	1942.9547	1924.9493	1925.9329	971.9799	N	518.2391	500.2286	501.2121	259.6196	4
18	2102.9905	2084.9800	2085.9634	1051.9952	C(+57.02)	404.1963	386.1856	387.1694	202.5981	3
19	2200.0432	2182.0327	2183.0161	1100.5216	P	244.1657	226.1549	227.1385	122.5828	2
20					K	147.1128	129.1022	130.0858	74.0564	1

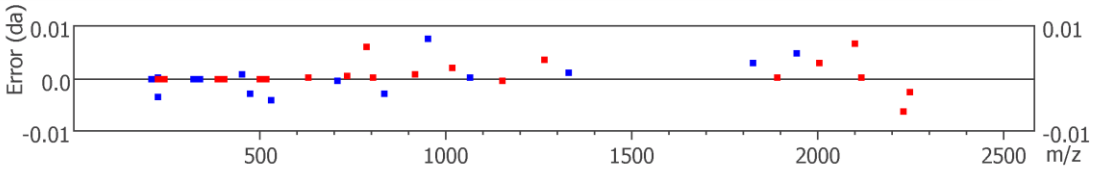
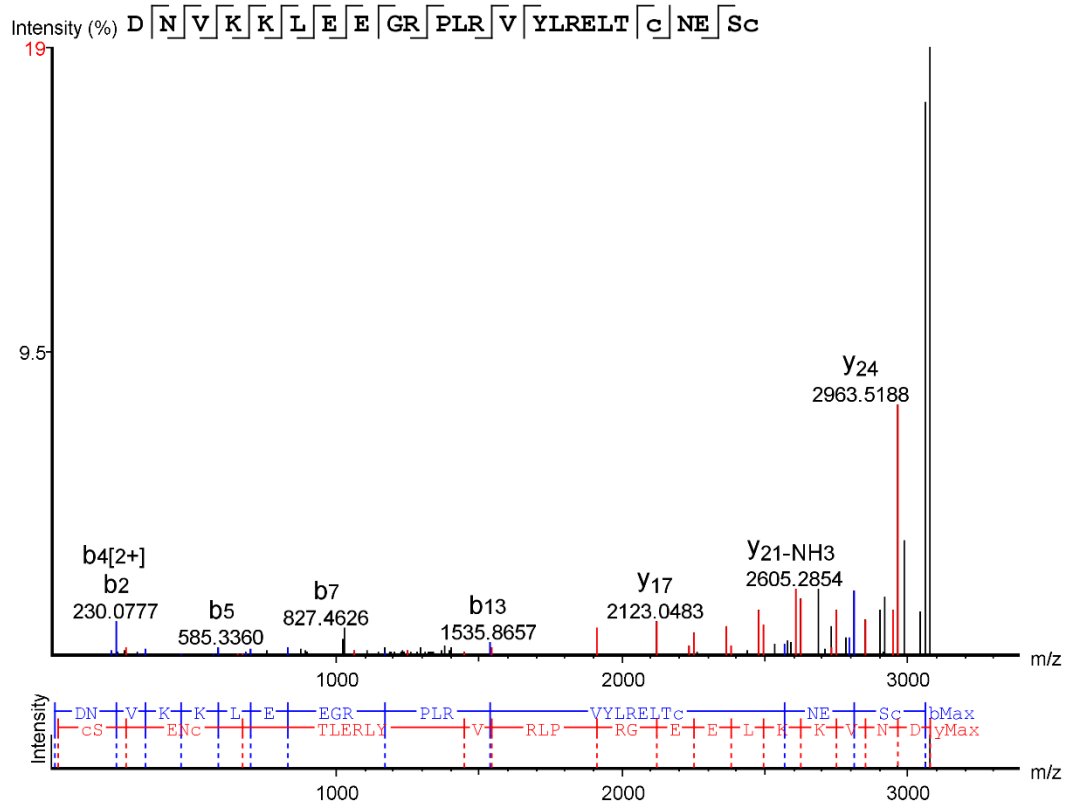


Figure 4.40 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* lys-C peptide l3.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					25
2	230.0773	212.0664	213.0505	115.5388	N	2963.5098	2945.5083	2946.4856	1482.2594	24
3	329.1454	311.1356	312.1191	165.0730	V	2849.4827	2831.4656	2832.4490	1425.2380	23
4	457.2425	439.2305	440.2141	229.1176	K	2750.4058	2732.3970	2733.3765	1375.7037	22
5	585.3372	567.3248	568.3079	293.1680	K	2622.3125	2604.3020	2605.2871	1311.6562	21
6	698.4188	680.4128	681.3931	349.7100	L	2494.2222	2476.2070	2477.1855	1247.6150	20
7	827.4642	809.4521	810.4357	414.2313	E	2381.1318	2363.1230	2364.1050	1191.0668	19
8	956.5052	938.4947	939.4783	478.7526	E	2252.0947	2234.0806	2235.0659	1126.5455	18
9	1013.5267	995.5162	996.4998	507.2634	G	2123.0425	2105.0378	2106.0212	1062.0251	17
10	1169.6321	1151.6173	1152.6008	585.3139	R	2066.0271	2048.0164	2049.0000	1033.5135	16
11	1266.6805	1248.6700	1249.6536	633.8403	P	1909.9355	1891.9153	1892.8988	955.4629	15
12	1379.7646	1361.7542	1362.7377	690.3823	L	1812.8730	1794.8625	1795.8461	906.9365	14
13	1535.8617	1517.8552	1518.8387	768.4329	R	1699.7891	1681.7786	1682.7621	850.3945	13
14	1634.9341	1616.9236	1617.9071	817.9670	V	1543.6913	1525.6774	1526.6609	772.3439	12
15	1797.9974	1779.9869	1780.9705	899.4987	Y	1444.6112	1426.6090	1427.5925	722.8098	11
16	1911.0815	1893.0710	1894.0546	956.0408	L	1281.5562	1263.5457	1264.5292	641.2781	10
17	2067.1826	2049.1721	2050.1555	1034.0913	R	1168.4722	1150.4617	1151.4452	584.7361	9
18	2196.2251	2178.2146	2179.1980	1098.6125	E	1012.3710	994.3605	995.3441	506.6855	8
19	2309.3093	2291.2988	2292.2822	1155.1547	L	883.3284	865.3179	866.3015	442.1642	7
20	2410.3569	2392.3464	2393.3298	1205.6785	T	770.2444	752.2338	753.2174	385.6222	6
21	2570.3843	2552.3772	2553.3606	1285.6938	C(+57.02)	669.1979	651.1861	652.1678	335.0983	5
22	2684.4307	2666.4202	2667.4036	1342.7153	N	509.1660	491.1555	492.1390	255.0830	4
23	2813.4663	2795.4626	2796.4385	1407.2366	E	395.1231	377.1125	378.0961	198.0616	3
24	2900.5051	2882.4946	2883.4780	1450.7526	S	266.0803	248.0700	249.0535	133.5403	2
25					C(+57.02)	179.0485	161.0379	162.0215	90.0242	1

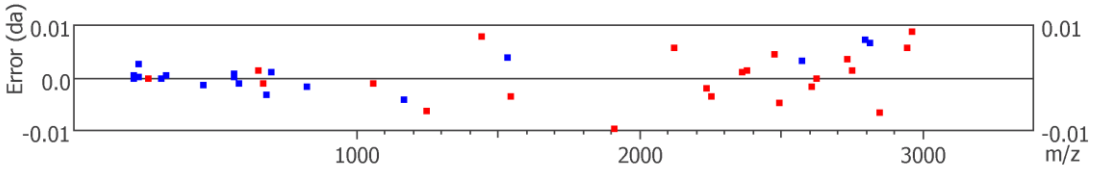
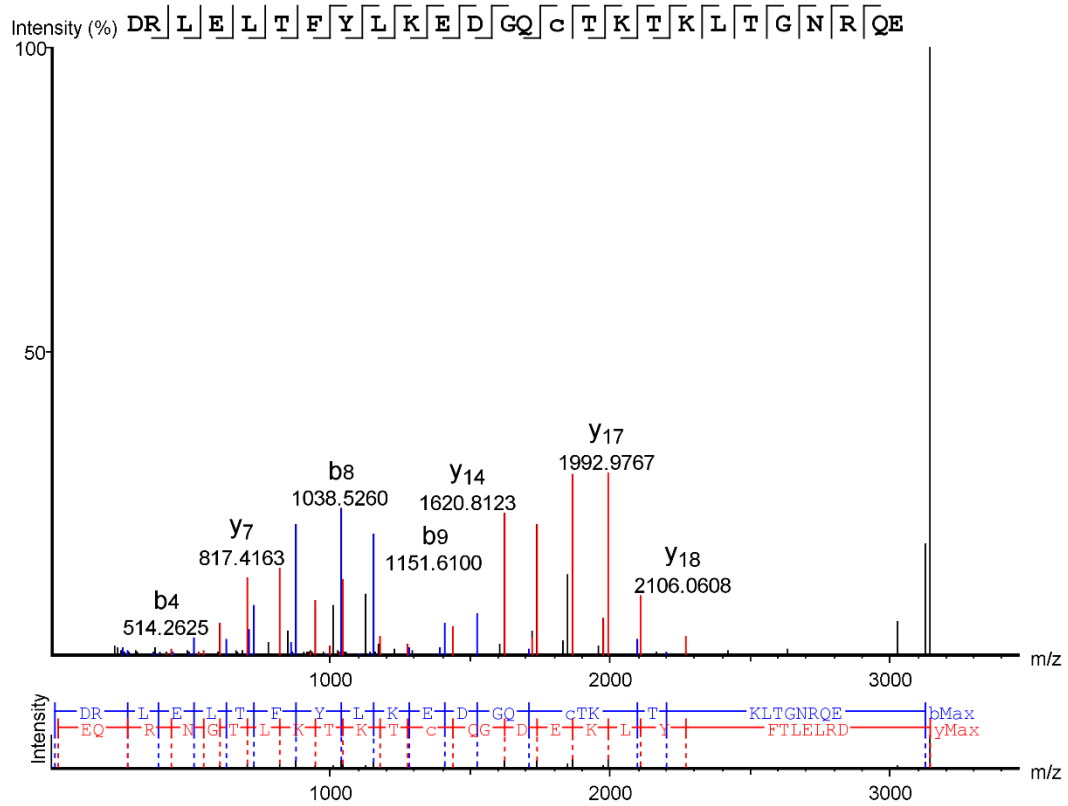


Figure 4.41 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* asp-N peptide a1.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					26
2	272.1349	254.1253	255.1085	136.5679	R	3028.5520	3010.5415	3011.5249	1514.7760	25
3	385.2182	367.2094	368.1929	193.1100	L	2872.4509	2854.4404	2855.4238	1436.7255	24
4	514.2621	496.2520	497.2347	257.6313	E	2759.3667	2741.3562	2742.3396	1380.1833	23
5	627.3459	609.3360	610.3196	314.1733	L	2630.3242	2612.3137	2613.2971	1315.6621	22
6	728.3944	710.3823	711.3672	364.6971	T	2517.2402	2499.2297	2500.2131	1259.1201	21
7	875.4615	857.4514	858.4357	438.2346	F	2416.1924	2398.1819	2399.1653	1208.5962	20
8	1038.5229	1020.5154	1021.4990	519.7630	Y	2269.1257	2251.1135	2252.0969	1135.0620	19
9	1151.6072	1133.5974	1134.5833	576.3050	L	2106.0569	2088.0503	2089.0337	1053.5304	18
10	1279.7074	1261.6945	1262.6780	640.3525	K	1992.9727	1974.9589	1975.9497	996.9902	17
11	1408.7465	1390.7305	1391.7206	704.8738	E	1864.8765	1846.8712	1847.8547	932.9409	16
12	1523.7755	1505.7640	1506.7476	762.3873	D	1735.8367	1717.8286	1718.8215	868.4196	15
13	1580.7960	1562.7855	1563.7690	790.8980	G	1620.8151	1602.8018	1603.7853	810.9061	14
14	1708.8521	1690.8441	1691.8276	854.9273	Q	1563.7908	1545.7803	1546.7638	782.3954	13
15	1868.8853	1850.8748	1851.8583	934.9426	C(+57.02)	1435.7305	1417.7217	1418.7052	718.3661	12
16	1969.9329	1951.9224	1952.9059	985.4756	T	1275.7076	1257.6910	1258.6746	638.3508	11
17	2098.0278	2080.0173	2081.0007	1049.5139	K	1174.6543	1156.6433	1157.6268	587.8269	10
18	2199.0847	2181.0649	2182.0483	1100.0463	T	1046.5588	1028.5483	1029.5388	523.7794	9
19	2327.1704	2309.1599	2310.1433	1164.0852	K	945.5112	927.5006	928.4887	473.2556	8
20	2440.2546	2422.2441	2423.2275	1220.6273	L	817.4150	799.4057	800.3893	409.2081	7
21	2541.3022	2523.2917	2524.2751	1271.1511	T	704.3318	686.3216	687.3052	352.6661	6
22	2598.3237	2580.3132	2581.2966	1299.6619	G	603.2844	585.2739	586.2575	302.1422	5
23	2712.3667	2694.3562	2695.3396	1356.6833	N	546.2653	528.2524	529.2368	273.6315	4
24	2868.4678	2850.4573	2851.4407	1434.7339	R	432.2199	414.2095	415.1949	216.6100	3
25	2996.5264	2978.5159	2979.4993	1498.7632	Q	276.1191	258.1087	259.0916	138.5595	2
26					E	148.0604	130.0499	131.0334	74.5302	1

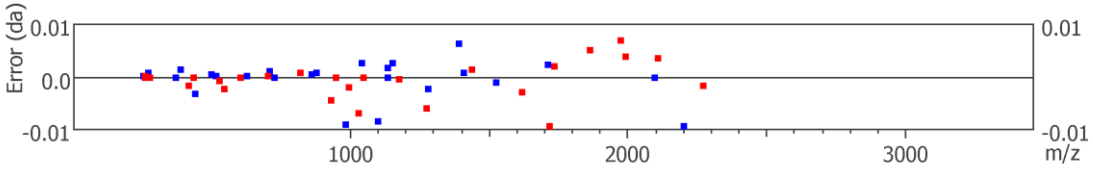
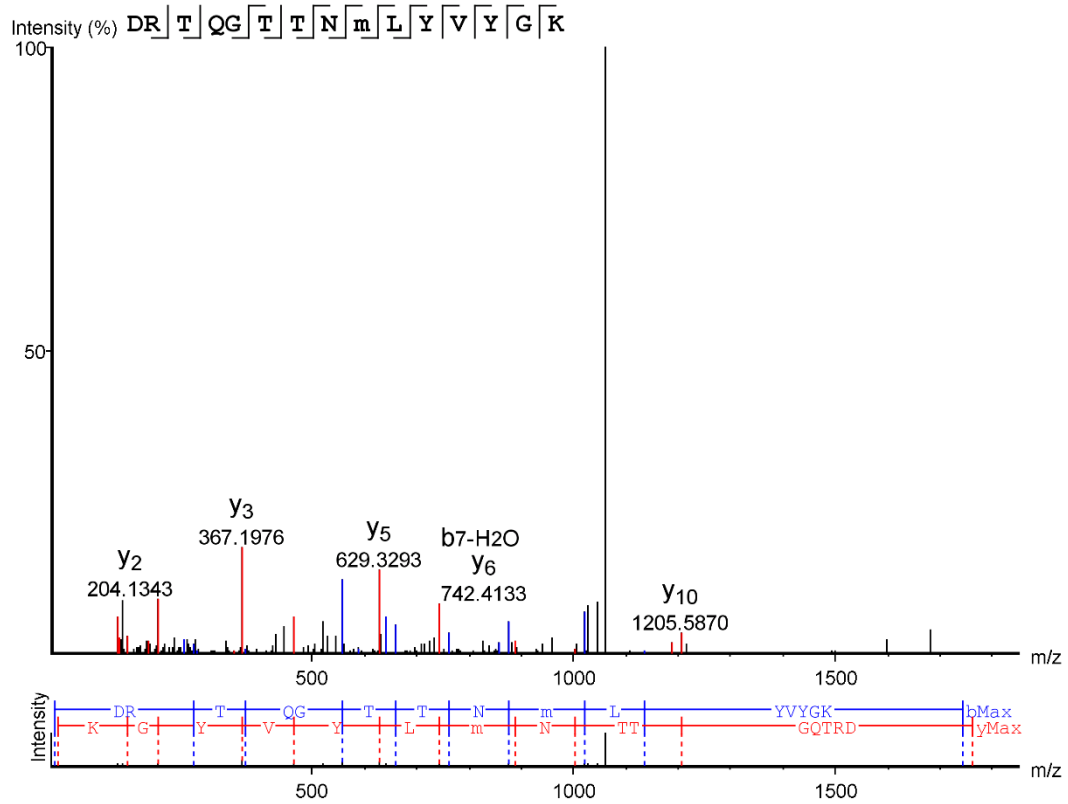


Figure 4.42 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* asp-N peptide a2.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					15
2	272.1352	254.1253	255.1086	136.5679	R	1647.8159	1629.8054	1630.7889	824.4080	14
3	373.1826	355.1730	356.1566	187.0918	T	1491.7148	1473.7043	1474.6879	746.3574	13
4	501.2421	483.2316	484.2151	251.1211	Q	1390.6671	1372.6566	1373.6401	695.8336	12
5	558.2627	540.2530	541.2366	279.6345	G	1262.6085	1244.5980	1245.5815	631.8043	11
6	659.3106	641.3008	642.2843	330.1556	T	1205.5935	1187.5765	1188.5636	603.2935	10
7	760.3596	742.3484	743.3320	380.6795	T	1104.5394	1086.5289	1087.5125	552.7697	9
8	874.4006	856.3895	857.3790	437.7009	N	1003.4967	985.4811	986.4647	502.2458	8
9	1021.4357	1003.4268	1004.4103	511.2187	M(+15....	889.4453	871.4382	872.4218	445.2244	7
10	1134.5190	1116.5109	1117.4944	567.7607	L	742.4141	724.4028	725.3864	371.7067	6
11	1297.5847	1279.5742	1280.5577	649.2924	Y	629.3293	611.3187	612.3023	315.1646	5
12	1396.6531	1378.6426	1379.6261	698.8265	V	466.2659	448.2554	449.2390	233.6330	4
13	1559.7164	1541.7059	1542.6895	780.3582	Y	367.1978	349.1869	350.1706	184.0988	3
14	1616.7379	1598.7274	1599.7109	808.8690	G	204.1345	186.1236	187.1076	102.5671	2
15					K	147.1128	129.1024	130.0863	74.0564	1

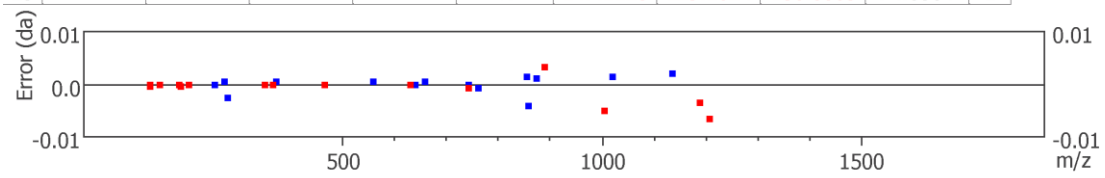
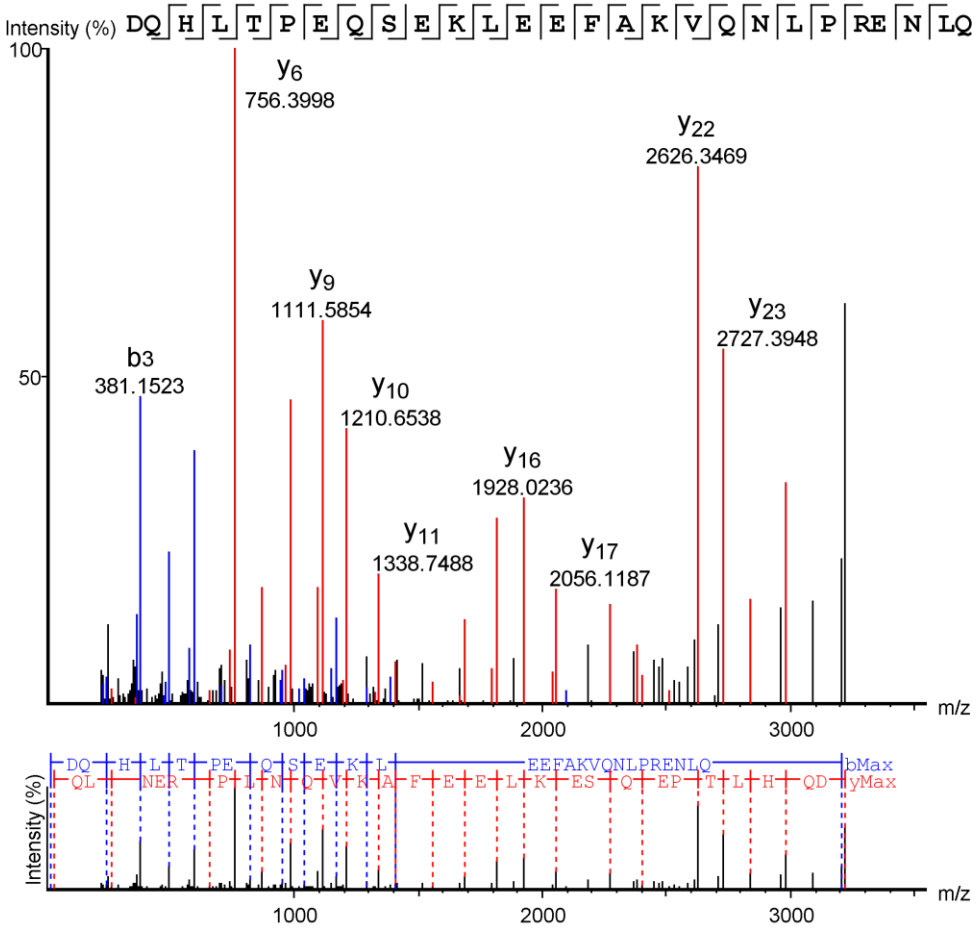


Figure 4.43 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* asp-N peptide a3.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					27
2	244.0926	226.0824	227.0660	122.5467	Q	3105.5962	3087.5857	3088.5691	1553.2981	26
3	381.1512	363.1411	364.1248	191.0761	H	2977.5330	2959.5271	2960.5105	1489.2688	25
4	494.2358	476.2244	477.2062	247.6181	L	2840.4756	2822.4683	2823.4517	1420.7394	24
5	595.2834	577.2725	578.2570	298.1420	T	2727.3936	2709.3843	2710.3677	1364.1974	23
6	692.3367	674.3262	675.3098	346.6684	P	2626.3433	2608.3364	2609.3198	1313.6735	22
7	821.3788	803.3688	804.3524	411.1897	E	2529.2942	2511.2891	2512.2671	1265.1471	21
8	949.4296	931.4274	932.4109	475.2214	Q	2400.2507	2382.2412	2383.2190	1200.6259	20
9	1036.4719	1018.4586	1019.4430	518.7350	S	2272.1931	2254.1826	2255.1660	1136.5966	19
10	1165.5120	1147.5010	1148.4856	583.2563	E	2185.1611	2167.1506	2168.1340	1093.0806	18
11	1293.6028	1275.5969	1276.5804	647.3037	K	2056.1125	2038.1080	2039.0979	1028.5593	17
12	1406.6824	1388.6810	1389.6604	703.8489	L	1928.0195	1910.0131	1910.9966	964.5118	16
13	1535.7341	1517.7236	1518.7072	768.3671	E	1814.9370	1796.9272	1797.9125	907.9697	15
14	1664.7767	1646.7662	1647.7498	832.8884	E	1685.9004	1667.8864	1668.8727	843.4484	14
15	1811.8451	1793.8346	1794.8181	906.4225	F	1556.8577	1538.8438	1539.8273	778.9271	13
16	1882.8822	1864.8717	1865.8552	941.9446	A	1409.7881	1391.7754	1392.7589	705.3929	12
17	2010.9772	1992.9667	1993.9502	1005.9886	K	1338.7540	1320.7383	1321.7178	669.8744	11
18	2110.0457	2092.0352	2093.0186	1055.5228	V	1210.6543	1192.6350	1193.6257	605.8269	10
19	2238.1042	2220.0938	2221.0771	1119.5521	Q	1111.5847	1093.5729	1094.5605	556.2927	9
20	2352.1472	2334.1367	2335.1201	1176.5736	N	983.5264	965.5163	966.5061	492.2634	8
21	2465.2312	2447.2207	2448.2041	1233.1156	L	869.4845	851.4733	852.4569	435.2419	7
22	2562.2839	2544.2734	2545.2568	1281.6420	P	756.3992	738.3893	739.3731	378.6999	6
23	2718.3850	2700.3745	2701.3579	1359.6925	R	659.3477	641.3365	642.3201	330.1736	5
24	2847.4277	2829.4172	2830.4006	1424.2139	E	503.2460	485.2354	486.2190	252.1230	4
25	2961.4705	2943.4600	2944.4434	1481.2352	N	374.2034	356.1926	357.1764	187.6017	3
26	3074.5547	3056.5442	3057.5276	1537.7773	L	260.1605	242.1499	243.1336	130.5802	2
27					Q	147.0764	129.0659	130.0494	74.0382	1

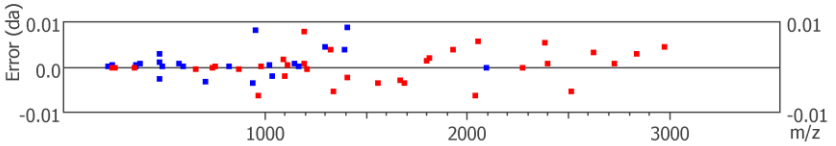
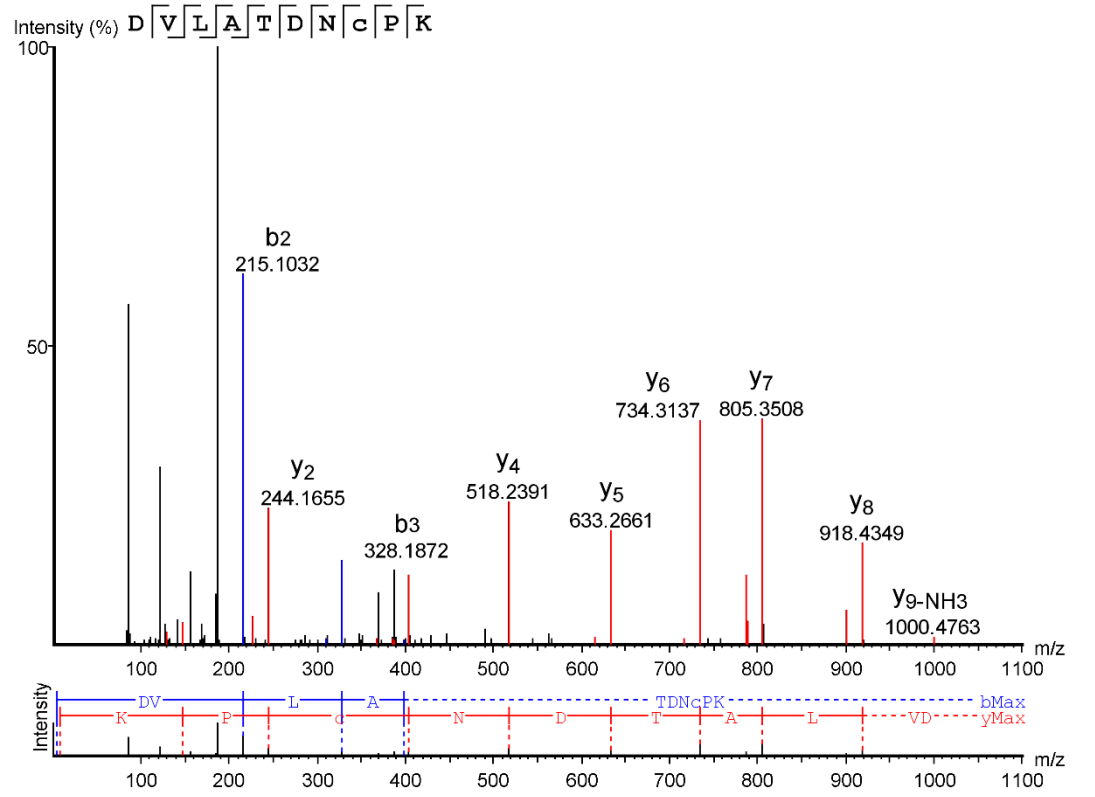


Figure 4.44 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* asp-N peptide a4.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					10
2	215.1024	197.0926	198.0762	108.0516	V	1017.5033	999.4927	1000.4825	509.2516	9
3	328.1865	310.1743	311.1602	164.5936	L	918.4368	900.4224	901.4080	459.7175	8
4	399.2241	381.2138	382.1974	200.1122	A	805.3502	787.3414	788.3338	403.1754	7
5	500.2720	482.2615	483.2451	250.6360	T	734.3137	716.3046	717.2867	367.6602	6
6	615.2990	597.2884	598.2720	308.1495	D	633.2663	615.2494	616.2391	317.1330	5
7	729.3419	711.3314	712.3149	365.1710	N	518.2392	500.2286	501.2122	259.6196	4
8	889.3726	871.3620	872.3456	445.1863	C(+57.02)	404.1962	386.1844	387.1677	202.5981	3
9	986.4253	968.4147	969.3983	493.7126	P	244.1653	226.1549	227.1385	122.5828	2
10					K	147.1127	129.1022	130.0865	74.0564	1

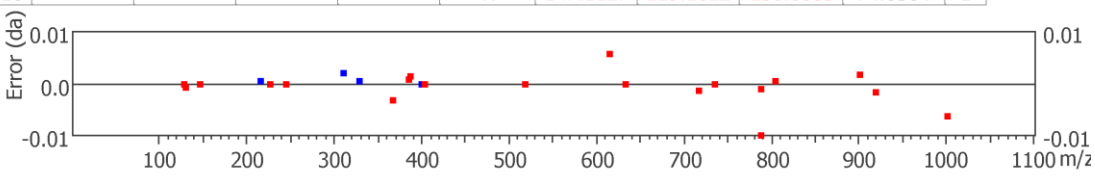
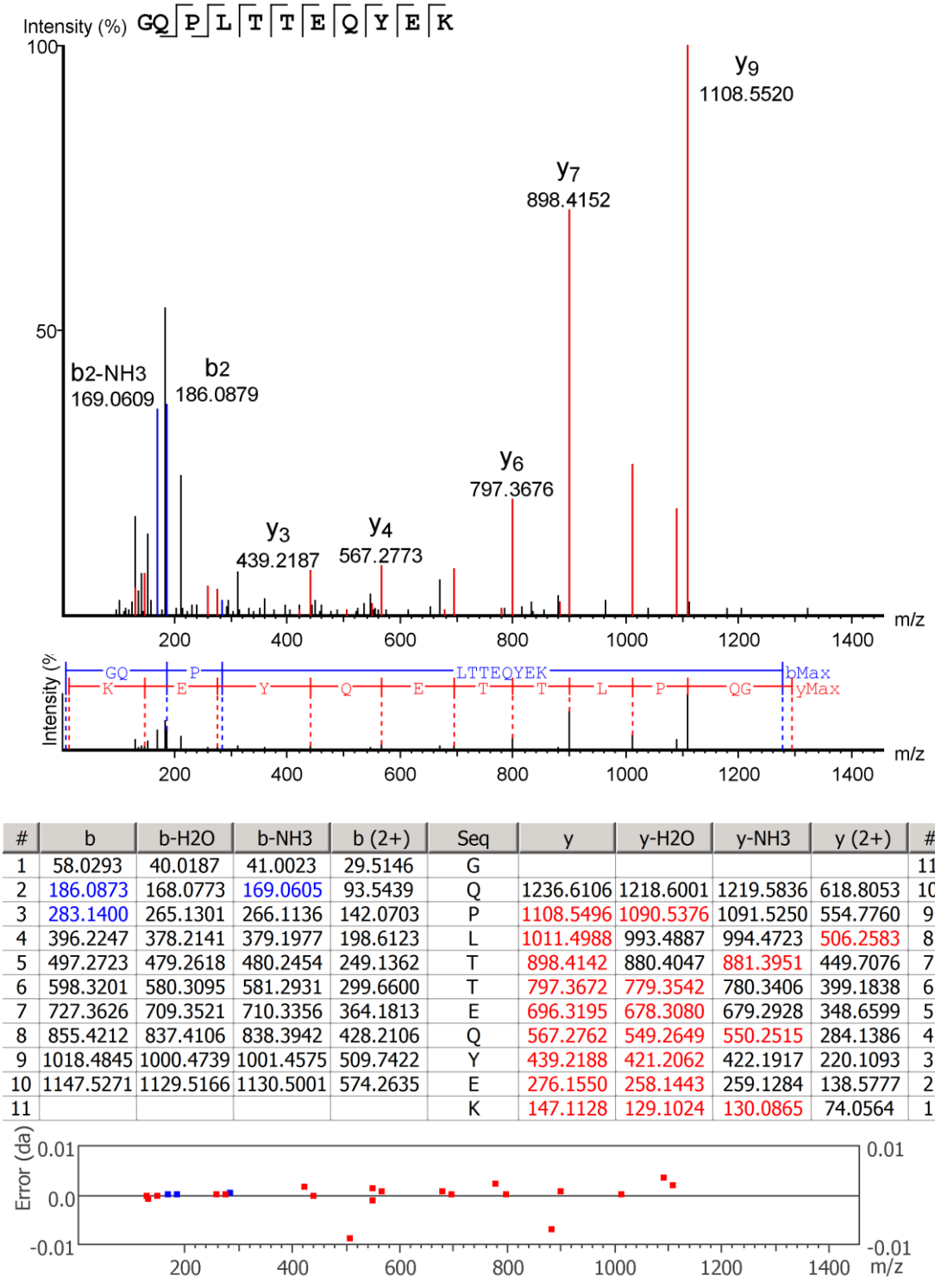


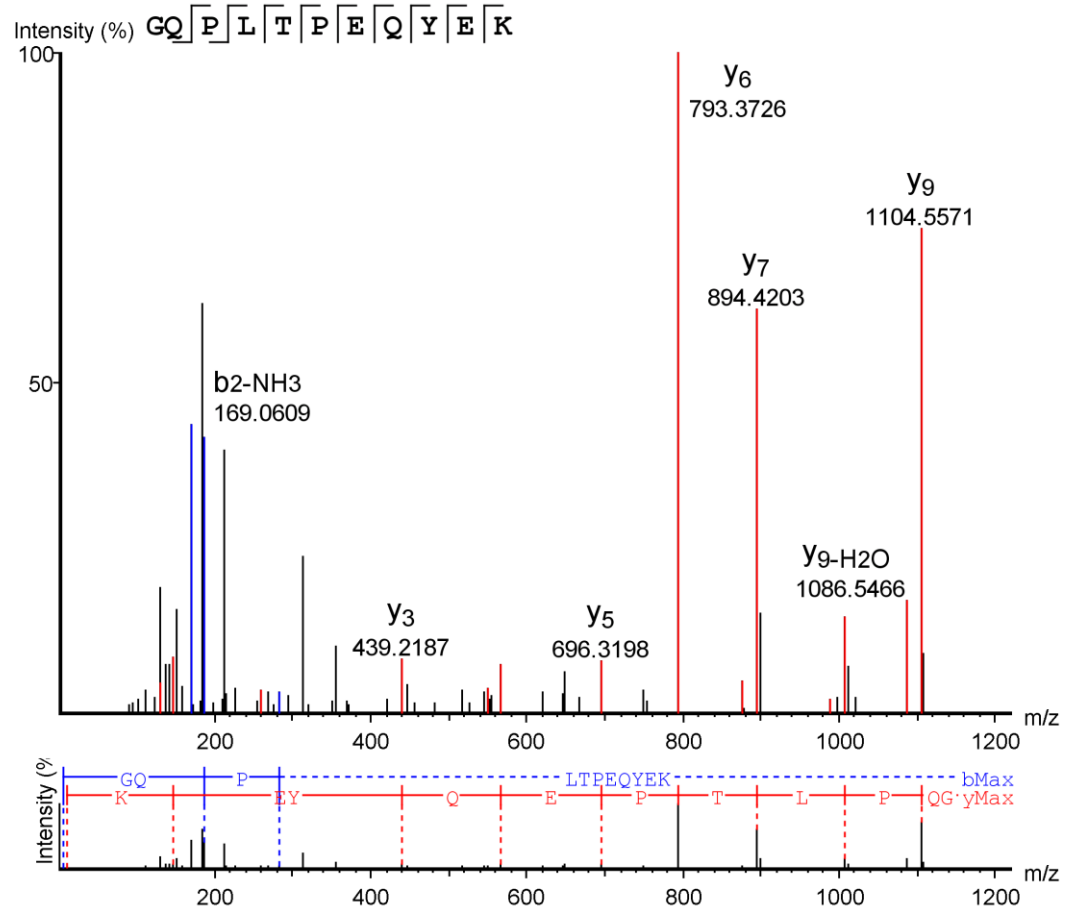
Figure 4.45 *De novo* sequence analysis of the processed MS/MS spectra of *M. agrestis* asp-N peptide a5.

4.6.2.2 Mutations



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					11
2	186.0873	168.0773	169.0605	93.5439	Q	1236.6106	1218.6001	1219.5836	618.8053	10
3	283.1400	265.1301	266.1136	142.0703	P	1108.5496	1090.5376	1091.5250	554.7760	9
4	396.2247	378.2141	379.1977	198.6123	L	1011.4988	993.4887	994.4723	506.2583	8
5	497.2723	479.2618	480.2454	249.1362	T	898.4142	880.4047	881.3951	449.7076	7
6	598.3201	580.3095	581.2931	299.6600	T	797.3672	779.3542	780.3406	399.1838	6
7	727.3626	709.3521	710.3356	364.1813	E	696.3195	678.3080	679.2928	348.6599	5
8	855.4212	837.4106	838.3942	428.2106	Q	567.2762	549.2649	550.2515	284.1386	4
9	1018.4845	1000.4739	1001.4575	509.7422	Y	439.2188	421.2062	422.1917	220.1093	3
10	1147.5271	1129.5166	1130.5001	574.2635	E	276.1550	258.1443	259.1284	138.5777	2
11					K	147.1128	129.1024	130.0865	74.0564	1

Figure 4.46 Evidence of a mutation in the processed MS/MS spectra of *M. agrestis* glareosin tryptic peptide t9.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					11
2	186.0872	168.0773	169.0607	93.5439	Q	1232.6157	1214.6052	1215.5887	616.8079	10
3	283.1396	265.1301	266.1136	142.0703	P	1104.5566	1086.5411	1087.5302	552.7786	9
4	396.2247	378.2141	379.1977	198.6123	L	1007.5004	989.4899	990.4774	504.2522	8
5	497.2723	479.2618	480.2454	249.1362	T	894.4197	876.4091	877.3933	447.7101	7
6	594.3251	576.3146	577.2982	297.6626	P	793.3724	775.3621	776.3456	397.1863	6
7	723.3677	705.3571	706.3407	362.1838	E	696.3168	678.3093	679.2928	348.6599	5
8	851.4263	833.4157	834.3993	426.2131	Q	567.2767	549.2661	550.2500	284.1386	4
9	1014.4896	996.4791	997.4626	507.7448	Y	439.2177	421.2081	422.1917	220.1093	3
10	1143.5322	1125.5217	1126.5052	572.2661	E	276.1554	258.1444	259.1284	138.5777	2
11					K	147.1129	129.1022	130.0862	74.0564	1

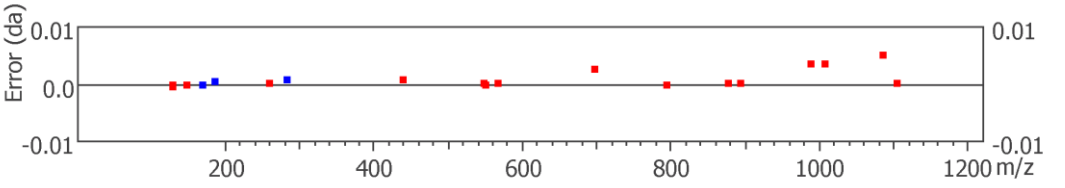
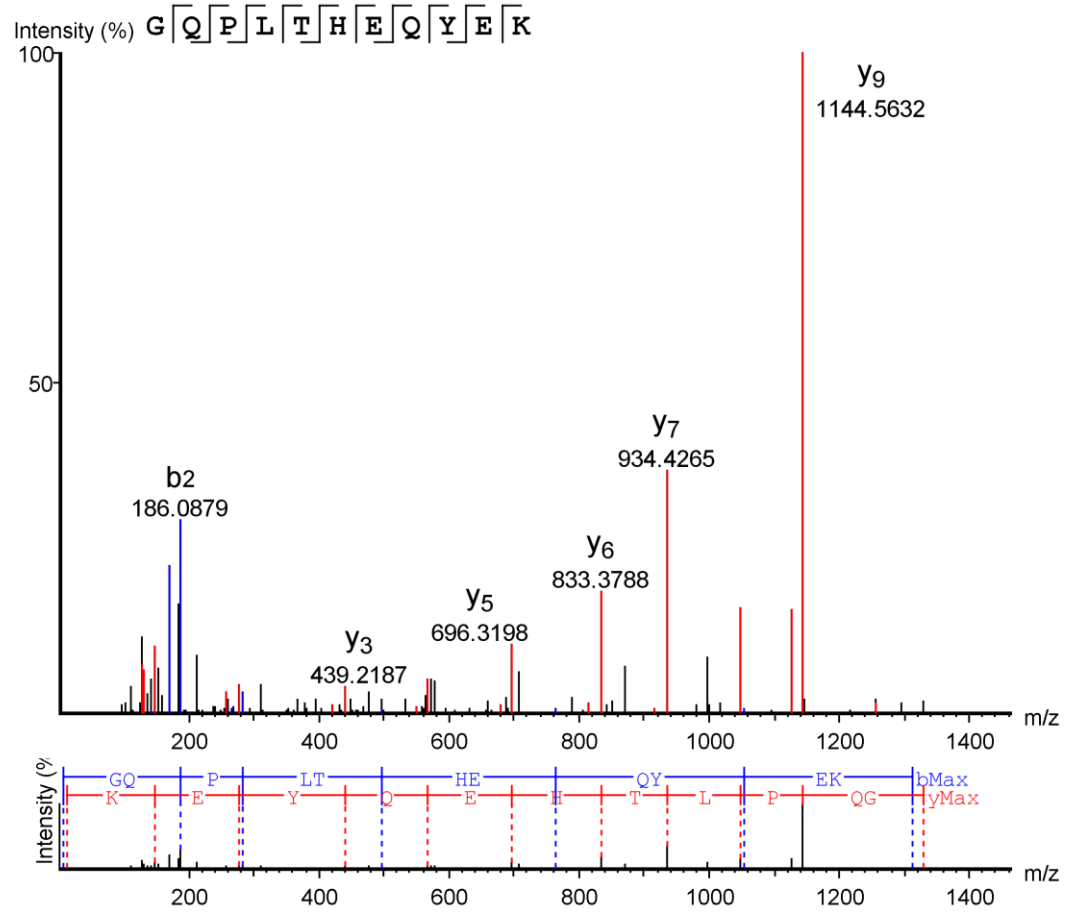


Figure 4.47 Evidence of a mutation in the processed MS/MS spectra of *M. agrestis* glareosin tryptic peptide t9.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					11
2	186.0872	168.0773	169.0609	93.5439	Q	1272.6218	1254.6113	1255.6014	636.8109	10
3	283.1403	265.1301	266.1131	142.0691	P	1144.5626	1126.5452	1127.5363	572.7816	9
4	396.2247	378.2141	379.1977	198.6123	L	1047.5084	1029.5000	1030.4835	524.2552	8
5	497.2740	479.2618	480.2454	249.1362	T	934.4252	916.4211	917.3995	467.7132	7
6	634.3313	616.3207	617.3043	317.6656	H	833.3780	815.3695	816.3518	417.1894	6
7	763.3727	745.3633	746.3469	382.1869	E	696.3177	678.3077	679.2928	348.6599	5
8	891.4324	873.4219	874.4055	446.2162	Q	567.2770	549.2675	550.2523	284.1386	4
9	1054.4930	1036.4852	1037.4688	527.7479	Y	439.2186	421.2128	422.1917	220.1093	3
10	1183.5383	1165.5278	1166.5114	592.2692	E	276.1553	258.1450	259.1284	138.5777	2
11					K	147.1128	129.1023	130.0863	74.0564	1

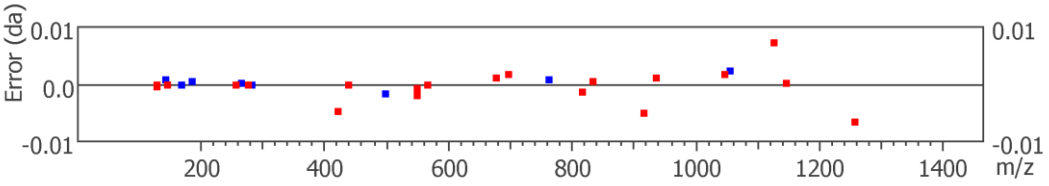
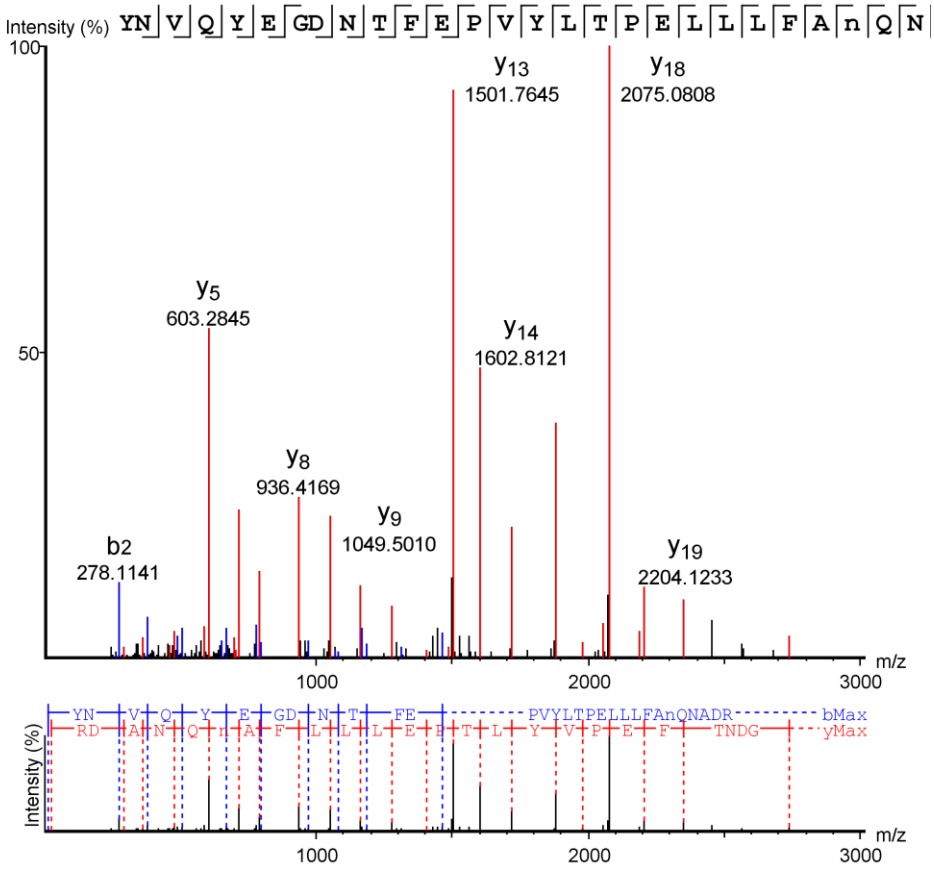


Figure 4.48 Evidence of a mutation in the processed MS/MS spectra of *M. agrestis* glareosin tryptic peptide t9.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	164.0712	146.0606	147.0441	82.5356	Y					30
2	278.1134	260.1035	261.0869	139.5570	N	3371.6064	3353.5959	3354.5793	1686.3032	29
3	377.1809	359.1719	360.1555	189.0912	V	3257.5637	3239.5532	3240.5366	1629.2819	28
4	505.2405	487.2305	488.2143	253.1205	Q	3158.4951	3140.4846	3141.4680	1579.7476	27
5	668.3033	650.2938	651.2756	334.6522	Y	3030.4365	3012.4260	3013.4094	1515.7183	26
6	797.3426	779.3363	780.3218	399.1735	E	2867.3733	2849.3628	2850.3462	1434.1866	25
7	854.3684	836.3578	837.3414	427.6842	G	2738.3342	2720.3203	2721.3037	1369.6654	24
8	969.3926	951.3848	952.3684	485.1977	D	2681.3093	2663.2988	2664.2822	1341.1547	23
9	1083.4388	1065.4279	1066.4187	542.2192	N	2566.2822	2548.2717	2549.2551	1283.6411	22
10	1184.4814	1166.4700	1167.4597	592.7430	T	2452.2395	2434.2290	2435.2124	1226.6198	21
11	1331.5544	1313.5439	1314.5267	666.2772	F	2351.1968	2333.1812	2334.1646	1176.0958	20
12	1460.5930	1442.5864	1443.5699	730.7985	E	2204.1245	2186.1133	2187.0962	1102.5616	19
13	1557.6498	1539.6393	1540.6228	779.3249	P	2075.0830	2057.0649	2058.0537	1038.0404	18
14	1656.7181	1638.7076	1639.6912	828.8591	V	1978.0331	1960.0175	1961.0010	989.5140	17
15	1819.7815	1801.7710	1802.7545	910.3907	Y	1878.9564	1860.9491	1861.9326	939.9798	16
16	1932.8655	1914.8550	1915.8385	966.9327	L	1715.8971	1697.8857	1698.8693	858.4481	15
17	2033.9132	2015.9027	2016.8862	1017.4566	T	1602.8137	1584.8016	1585.7852	801.9061	14
18	2130.9661	2112.9556	2113.9390	1065.9830	P	1501.7688	1483.7540	1484.7429	751.3823	13
19	2260.0085	2241.9980	2242.9814	1130.5043	E	1404.7086	1386.7012	1387.6847	702.8558	12
20	2373.0925	2355.0820	2356.0654	1187.0463	L	1275.6733	1257.6587	1258.6422	638.3346	11
21	2486.1768	2468.1663	2469.1497	1243.5884	L	1162.5859	1144.5746	1145.5581	581.7925	10
22	2599.2607	2581.2502	2582.2336	1300.1304	L	1049.4977	1031.4905	1032.4740	525.2505	9
23	2746.3291	2728.3186	2729.3020	1373.6646	F	936.4154	918.4064	919.3900	468.7085	8
24	2817.3662	2799.3557	2800.3391	1409.1831	A	789.3486	771.3380	772.3215	395.1743	7
25	2932.3933	2914.3828	2915.3662	1466.6967	N(+.98)	718.3115	700.2986	701.2871	359.6557	6
26	3060.4517	3042.4412	3043.4246	1530.7258	Q	603.2840	585.2725	586.2588	302.1422	5
27	3174.4946	3156.4841	3157.4675	1587.7473	N	475.2257	457.2153	458.1995	238.1129	4
28	3245.5317	3227.5212	3228.5046	1623.2659	A	361.1825	343.1724	344.1560	181.0915	3
29	3360.5588	3342.5483	3343.5317	1680.7794	D	290.1456	272.1353	273.1189	145.5729	2
30					R	175.1189	157.1084	158.0919	88.0595	1

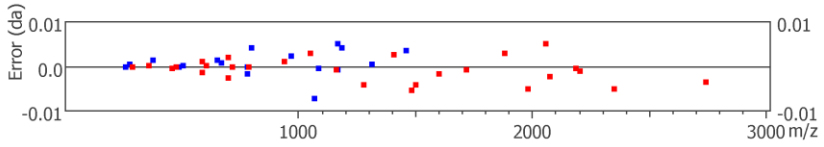
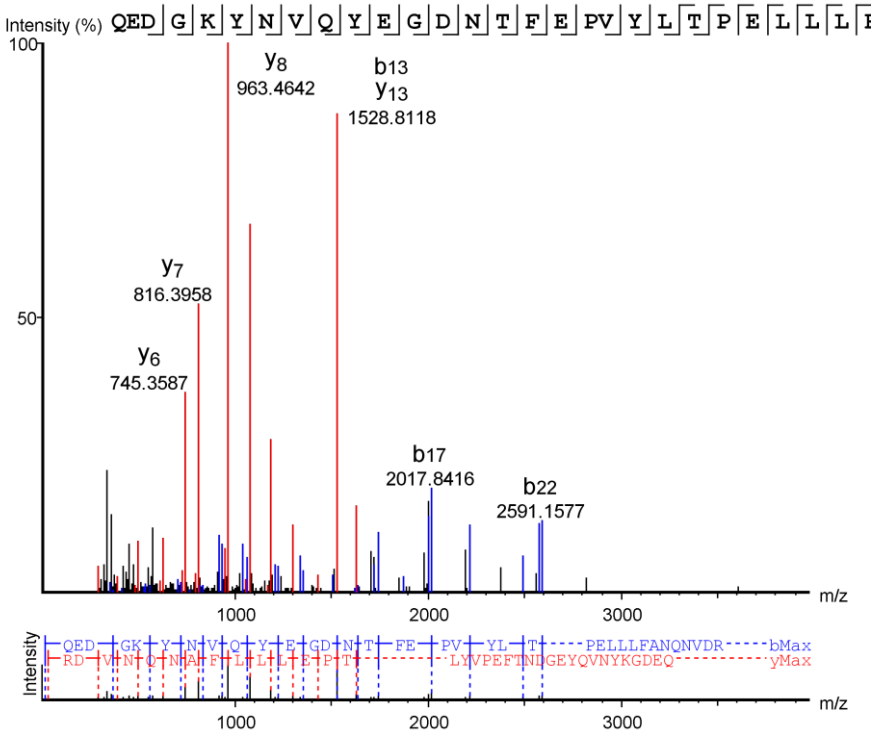


Figure 4.49 Evidence of a mutation in the processed MS/MS spectra of *M. agrestis* glareosin tryptic peptide t9.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	129.0664	111.0558	112.0394	65.0332	Q					35
2	258.1090	240.0984	241.0820	129.5545	E	3990.9031	3972.8926	3973.8760	1995.9515	34
3	373.1340	355.1249	356.1088	187.0680	D	3861.8606	3843.8501	3844.8335	1931.4303	33
4	430.1574	412.1459	413.1304	215.5787	G	3746.8335	3728.8230	3729.8064	1873.9167	32
5	558.2546	540.2406	541.2258	279.6262	K	3689.8120	3671.8015	3672.7849	1845.4060	31
6	721.3134	703.3049	704.2903	361.1578	Y	3561.7170	3543.7065	3544.6899	1781.3585	30
7	835.3573	817.3480	818.3330	418.1793	N	3398.6538	3380.6433	3381.6267	1699.8269	29
8	934.4265	916.4139	917.4000	467.7135	V	3284.6108	3266.6003	3267.5837	1642.8054	28
9	1062.4884	1044.4736	1045.4587	531.7460	Q	3185.5425	3167.5320	3168.5154	1593.2712	27
10	1225.5465	1207.5388	1208.5269	613.2745	Y	3057.4839	3039.4734	3040.4568	1529.2419	26
11	1354.5830	1336.5814	1337.5701	677.7958	E	2894.4207	2876.4102	2877.3936	1447.7103	25
12	1411.6129	1393.6024	1394.5859	706.3046	G	2765.3779	2747.3674	2748.3508	1383.1890	24
13	1526.6455	1508.6294	1509.6169	763.8265	D	2708.3564	2690.3459	2691.3293	1354.6782	23
14	1640.6740	1622.6724	1623.6609	820.8414	N	2593.3296	2575.3191	2576.3025	1297.1648	22
15	1741.7278	1723.7200	1724.7010	871.3652	T	2479.2866	2461.2761	2462.2595	1240.1433	21
16	1888.7990	1870.7885	1871.7773	944.8995	F	2378.2390	2360.2285	2361.2119	1189.6195	20
17	2017.8379	1999.8311	2000.8164	1009.4208	E	2231.1707	2213.1602	2214.1436	1116.0853	19
18	2114.8943	2096.8838	2097.8672	1057.9471	P	2102.1279	2084.1174	2085.1008	1051.5640	18
19	2213.9561	2195.9521	2196.9355	1107.4813	V	2005.0752	1987.0647	1988.0482	1003.0376	17
20	2377.0259	2359.0154	2359.9988	1189.0168	Y	1906.0068	1887.9963	1888.9799	953.5034	16
21	2490.1096	2472.0996	2473.0830	1245.5551	L	1742.9435	1724.9330	1725.9165	871.9717	15
22	2591.1499	2573.1436	2574.1306	1296.0789	T	1629.8585	1611.8490	1612.8325	815.4297	14
23	2688.2104	2670.2000	2671.1833	1344.6052	P	1528.8127	1510.7948	1511.7848	764.9059	13
24	2817.2532	2799.2427	2800.2261	1409.1266	E	1431.7583	1413.7485	1414.7321	716.3795	12
25	2930.3372	2912.3267	2913.3101	1465.6686	L	1302.7114	1284.7059	1285.6895	651.8582	11
26	3043.4211	3025.4106	3026.3940	1522.2106	L	1189.6313	1171.6218	1172.6091	595.3162	10
27	3156.5054	3138.4949	3139.4783	1578.7527	L	1076.5471	1058.5378	1059.5293	538.7742	9
28	3303.5737	3285.5632	3286.5466	1652.2869	F	963.4635	945.4537	946.4387	482.2321	8
29	3374.6108	3356.6003	3357.5837	1687.8054	A	816.3949	798.3853	799.3701	408.6979	7
30	3488.6538	3470.6433	3471.6267	1744.8269	N	745.3579	727.3481	728.3309	373.1794	6
31	3616.7124	3598.7019	3599.6853	1808.8562	Q	631.3157	613.3073	614.2906	316.1579	5
32	3730.7551	3712.7446	3713.7280	1865.8776	N	503.2569	485.2466	486.2305	252.1286	4
33	3829.8237	3811.8132	3812.7966	1915.4119	V	389.2140	371.2037	372.1873	195.1071	3
34	3944.8506	3926.8401	3927.8235	1972.9253	D	290.1456	272.1353	273.1189	145.5729	2
35					R	175.1189	157.1084	158.0919	88.0595	1

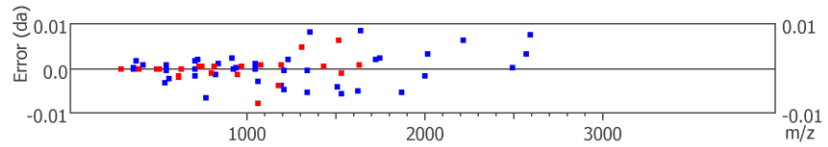
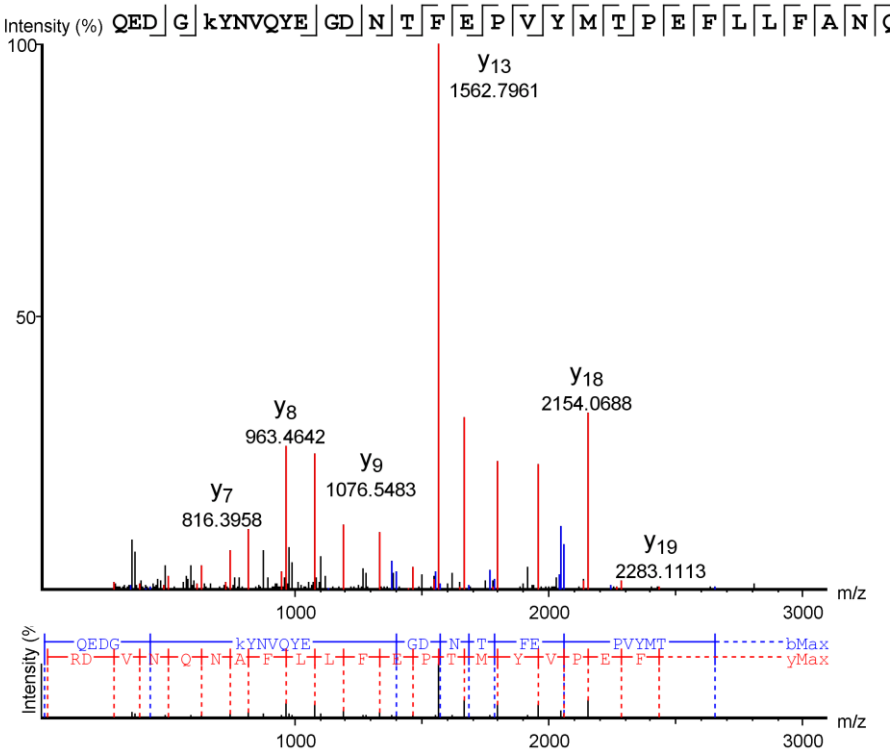


Figure 4.50 Evidence of a mutation in the processed MS/MS spectra of *M. agrestis* glareosin tryptic peptide t7.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	129.0664	111.0558	112.0394	65.0332	Q					35
2	258.1090	240.0984	241.0820	129.5545	E	4085.8496	4067.8391	4068.8225	2043.4248	34
3	373.1359	355.1246	356.1082	187.0680	D	3956.8071	3938.7966	3939.7800	1978.9036	33
4	430.1578	412.1486	413.1298	215.5787	G	3841.7803	3823.7698	3824.7532	1921.3901	32
5	601.2582	583.2476	584.2312	301.1291	K(+43.01)	3784.7588	3766.7483	3767.7317	1892.8794	31
6	764.3215	746.3109	747.2945	382.6607	Y	3613.6580	3595.6475	3596.6309	1807.3290	30
7	878.3644	860.3538	861.3374	439.6822	N	3450.5947	3432.5842	3433.5676	1725.7974	29
8	977.4328	959.4222	960.4058	489.2164	V	3336.5518	3318.5413	3319.5247	1668.7759	28
9	1105.4915	1087.4810	1088.4645	553.2457	Q	3237.4832	3219.4727	3220.4561	1619.2416	27
10	1268.5547	1250.5442	1251.5277	634.7773	Y	3109.4248	3091.4143	3092.3977	1555.2124	26
11	1397.5953	1379.5868	1380.5769	699.2986	E	2946.3613	2928.3508	2929.3342	1473.6807	25
12	1454.6188	1436.6083	1437.5918	727.8094	G	2817.3188	2799.3083	2800.2917	1409.1594	24
13	1569.6384	1551.6353	1552.6222	785.3229	D	2760.2974	2742.2869	2743.2703	1380.6487	23
14	1683.6832	1665.6781	1666.6565	842.3443	N	2645.2703	2627.2598	2628.2432	1323.1351	22
15	1784.7372	1766.7258	1767.7181	892.8682	T	2531.2275	2513.2170	2514.2004	1266.1138	21
16	1931.8047	1913.7942	1914.7777	966.4023	F	2430.1709	2412.1692	2413.1526	1215.5898	20
17	2060.8442	2042.8396	2043.8226	1030.9237	E	2283.1069	2265.1064	2266.0842	1142.0557	19
18	2157.9001	2139.8896	2140.8730	1079.4501	P	2154.0664	2136.0583	2137.0435	1077.5344	18
19	2256.9685	2238.9580	2239.9414	1128.9873	V	2057.0112	2039.0055	2039.9890	1029.0081	17
20	2420.0317	2402.0212	2403.0046	1210.5159	Y	1957.9460	1939.9371	1940.9207	979.4738	16
21	2551.0723	2533.0618	2534.0452	1276.0361	M	1794.8804	1776.8812	1777.8666	897.9421	15
22	2652.1274	2634.1094	2635.0928	1326.5599	T	1663.8406	1645.8361	1646.8168	832.4219	14
23	2749.1729	2731.1624	2732.1458	1375.0864	P	1562.7948	1544.7870	1545.7686	781.8981	13
24	2878.2153	2860.2048	2861.1882	1439.6077	E	1465.7405	1447.7388	1448.7164	733.3717	12
25	3025.2837	3007.2732	3008.2566	1513.1418	F	1336.6998	1318.6903	1319.6738	668.8504	11
26	3138.3677	3120.3572	3121.3406	1569.6838	L	1189.6321	1171.6218	1172.6053	595.3162	10
27	3251.4519	3233.4414	3234.4248	1626.2260	L	1076.5476	1058.5378	1059.5214	538.7742	9
28	3398.5203	3380.5098	3381.4932	1699.7601	F	963.4624	945.4537	946.4285	482.2321	8
29	3469.5574	3451.5469	3452.5303	1735.2787	A	816.3952	798.3853	799.3688	408.6979	7
30	3583.6003	3565.5898	3566.5732	1792.3002	N	745.3582	727.3481	728.3311	373.1794	6
31	3711.6589	3693.6484	3694.6318	1856.3295	Q	631.3154	613.3000	614.2875	316.1579	5
32	3825.7019	3807.6914	3808.6748	1913.3510	N	503.2568	485.2466	486.2302	252.1286	4
33	3924.7703	3906.7598	3907.7432	1962.8851	V	389.2148	371.2037	372.1873	195.1071	3
34	4039.7971	4021.7866	4022.7700	2020.3986	D	290.1460	272.1353	273.1189	145.5729	2
35					R	175.1189	157.1084	158.0919	88.0595	1

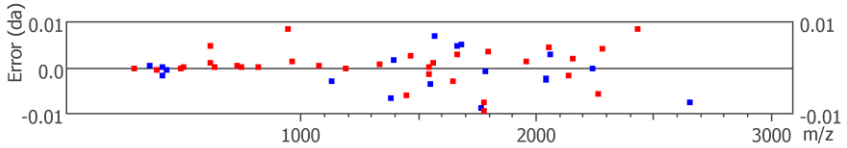


Figure 4.51 Evidence of a mutation in the processed MS/MS spectra of *M. agrestis* glareosin tryptic peptide t7.

4.6.2.3 *Peptide map of glareosin variants for individual males with distinctive protein profiles, used to confirm mutation sequencing.*

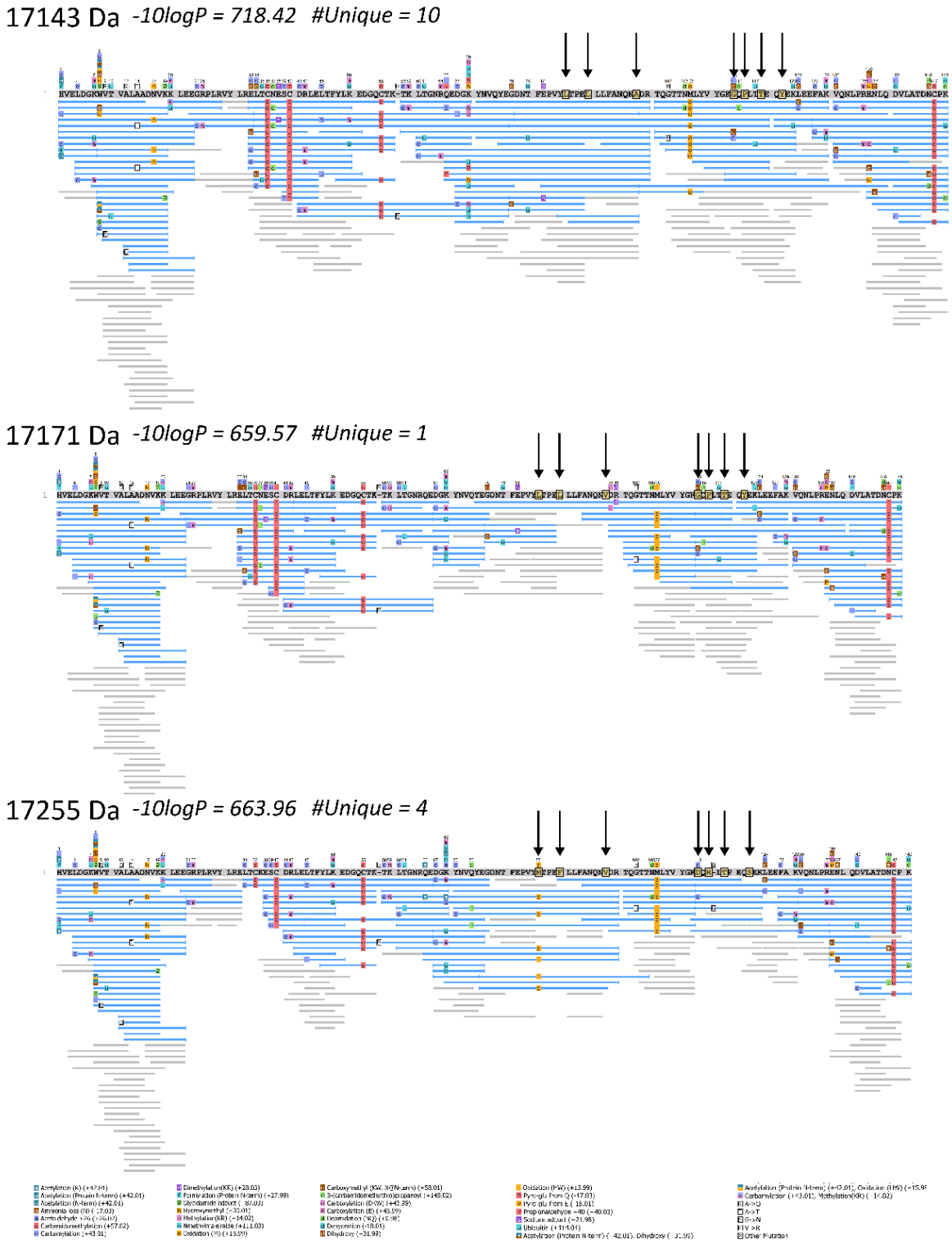


Figure 4.52 Sequence coverage of peptides identified from analysis of urinary proteins in Male 1, Sample A (Male 1, lane 1 from Figure 4.1), searched against four different variants of the novel *M. agrestis* glareosin protein using the SPIDER algorithm from PEAKS™. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

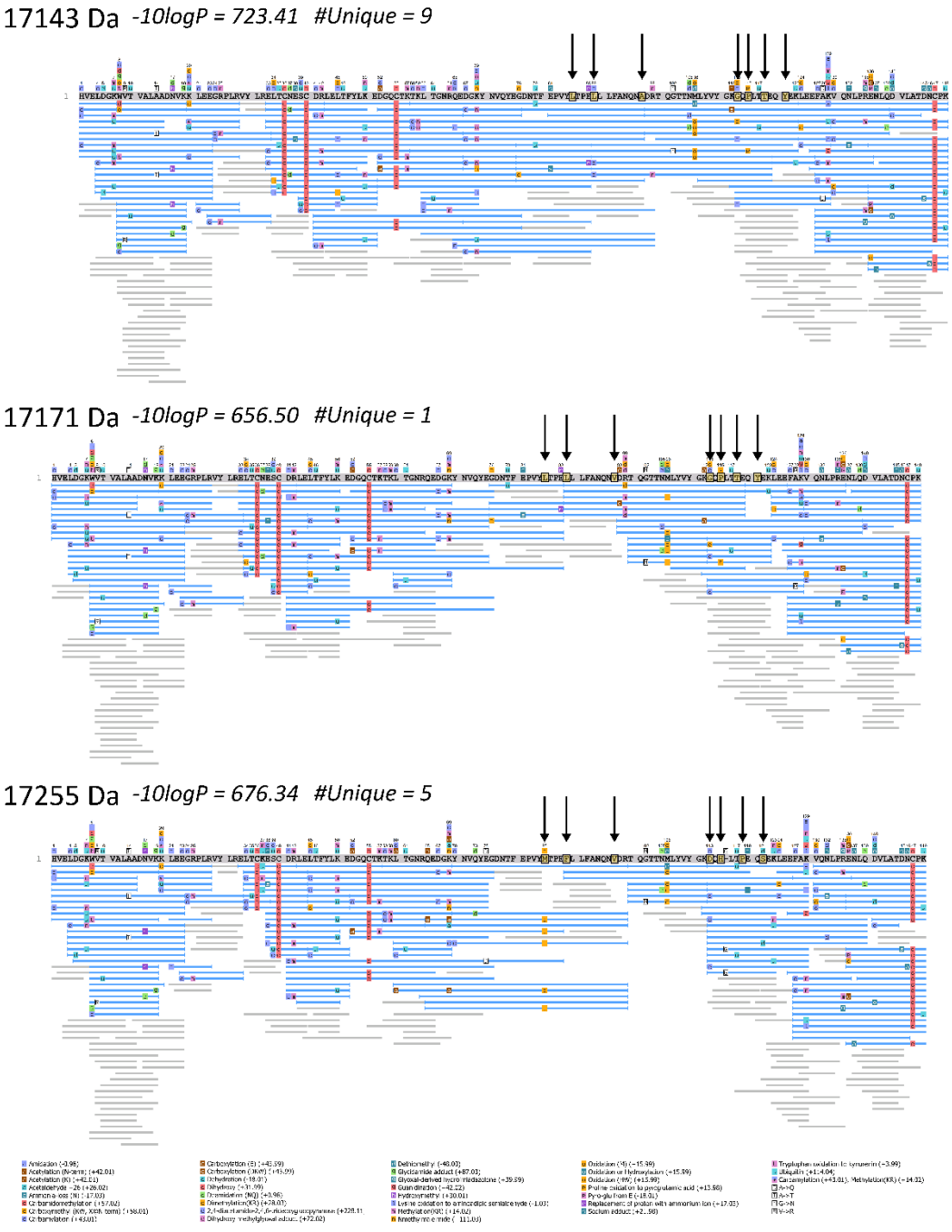


Figure 4.53 Sequence coverage of peptides identified from analysis of urinary proteins in Male 2, Sample A (Male 2, lane 1 from Figure 4.1), searched against four different variants of the novel *M. agrestis* glareosin protein using the SPIDER algorithm from PEAKS™. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

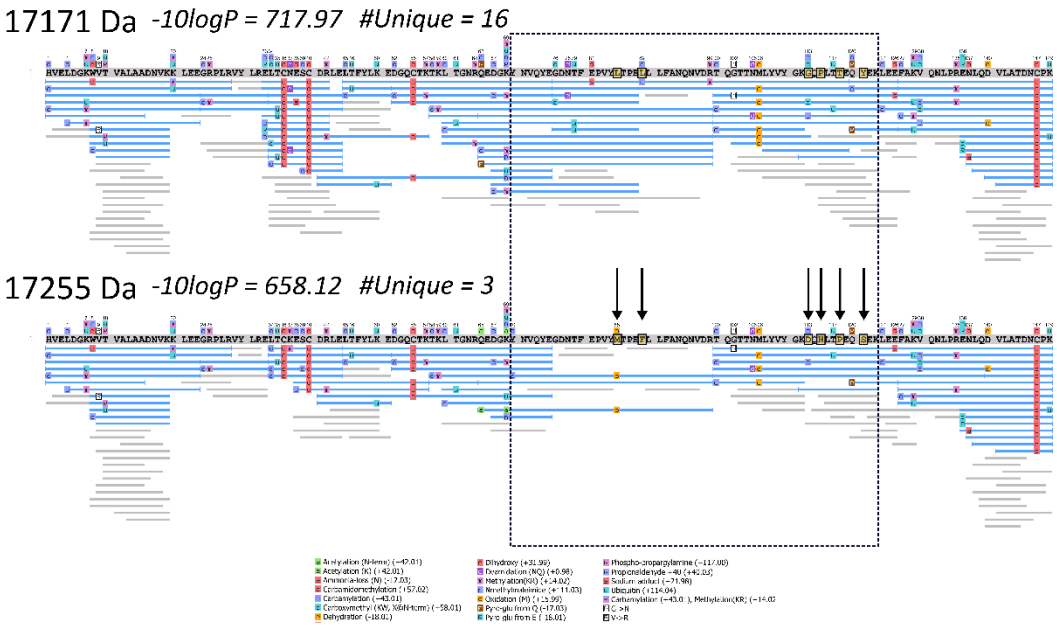


Figure 4.54 Sequence coverage of peptides identified from analysis of urinary proteins in Male 3, Sample A (Male 3, lane 1 from Figure 4.1), searched against four different variants of the novel *M. agrestis* glareosin protein using the SPIDER algorithm from PEAKS™. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

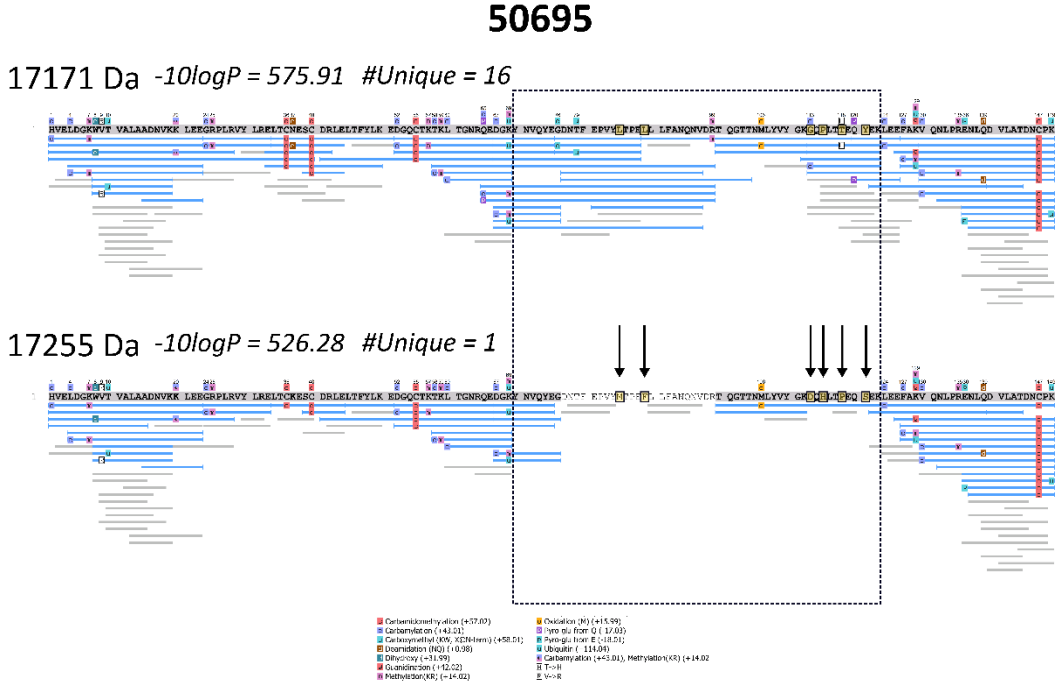


Figure 4.55 Sequence coverage of peptides identified from analysis of urinary proteins in Male 11, searched against four different variants of the novel *M. agrestis* glareosin protein using the SPIDER algorithm from PEAKS™. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

4.6.3 Sequence coverage of odorant binding proteins from analysis of pooled female field vole urine.

Figure 4.56 Sequence coverage of peptides matching OBP3 from *M. glareolus* from the analysis of proteins in pooled female urine, using the SPIDER algorithm from PEAKS^{TM}. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

Figure 4.57 Sequence coverage of peptides matching OBP2 from *M. glareolus* from the analysis of proteins in pooled female urine, using the SPIDER algorithm from PEAKS^{TM}. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

Figure 4.58 Sequence coverage of peptides matching OBP1 from *M. glareolus* from the analysis of proteins in pooled female urine, using the SPIDER algorithm from PEAKS^{TM}. Blue lines indicate an identified peptide, grey lines indicate a *de novo* tag.

4.6.4 Alignment of major urinary proteins to support preliminary sequencing of a field vole major urinary protein.

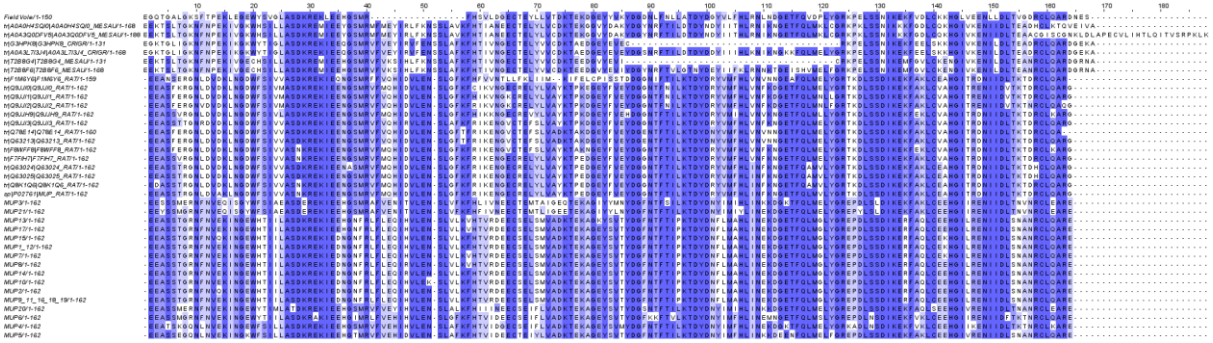


Figure 4.59 Multiple sequence alignment of major urinary proteins from *Rodentia*, generated in Clustal omega [https://www.ebi.ac.uk/Tools/msa/clustalo/] and processed in JalView [https://www.jalview.org/].

4.6.5 Intact mass profiles of juvenile field vole urine.

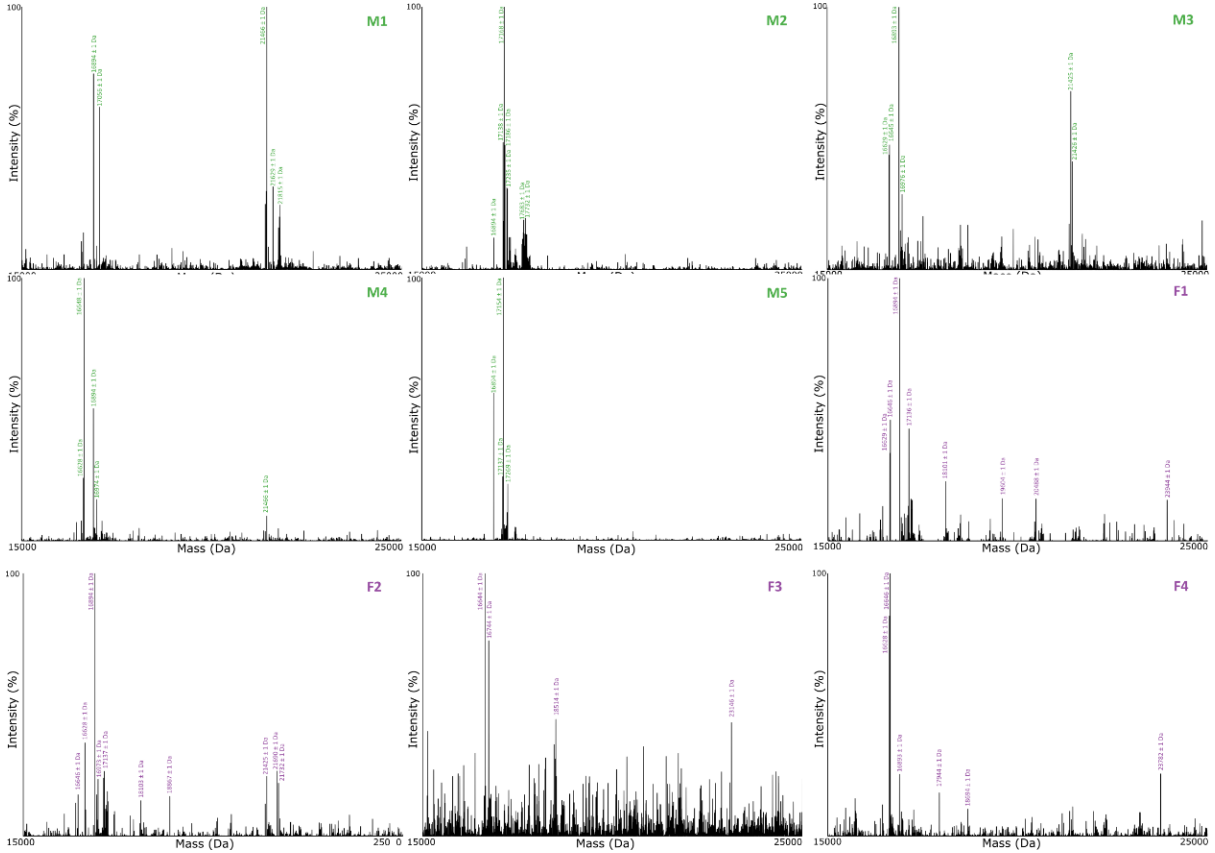


Figure 4.60 Deconvoluted mass spectra generated by analysis of intact proteins in the urine of juvenile field voles. Males, M1-5; Females F1-4.

5 Characterisation of the urinary protein content in the New Zealand brushtail possum, *Trichosurus vulpecula*

5.6 Supplementary Information

5.6.1 Intact mass profiles of individual samples

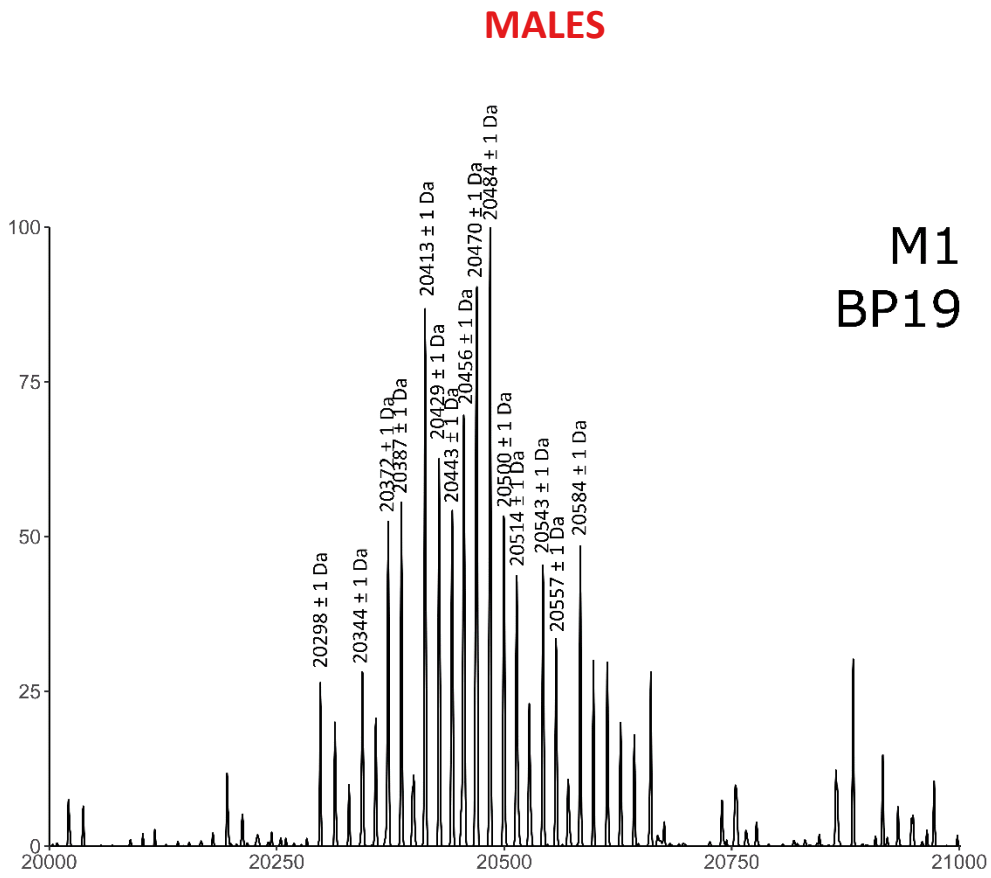


Figure 5.1 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 1 (male 1, sample BP19).

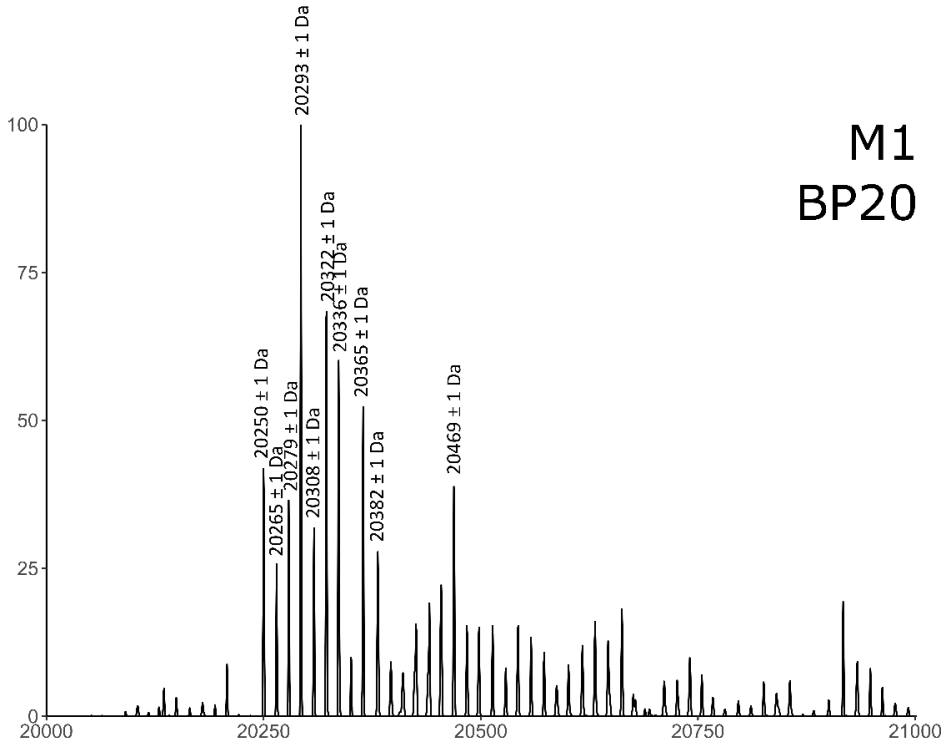


Figure 5.2 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 1 (male 1, sample BP20).

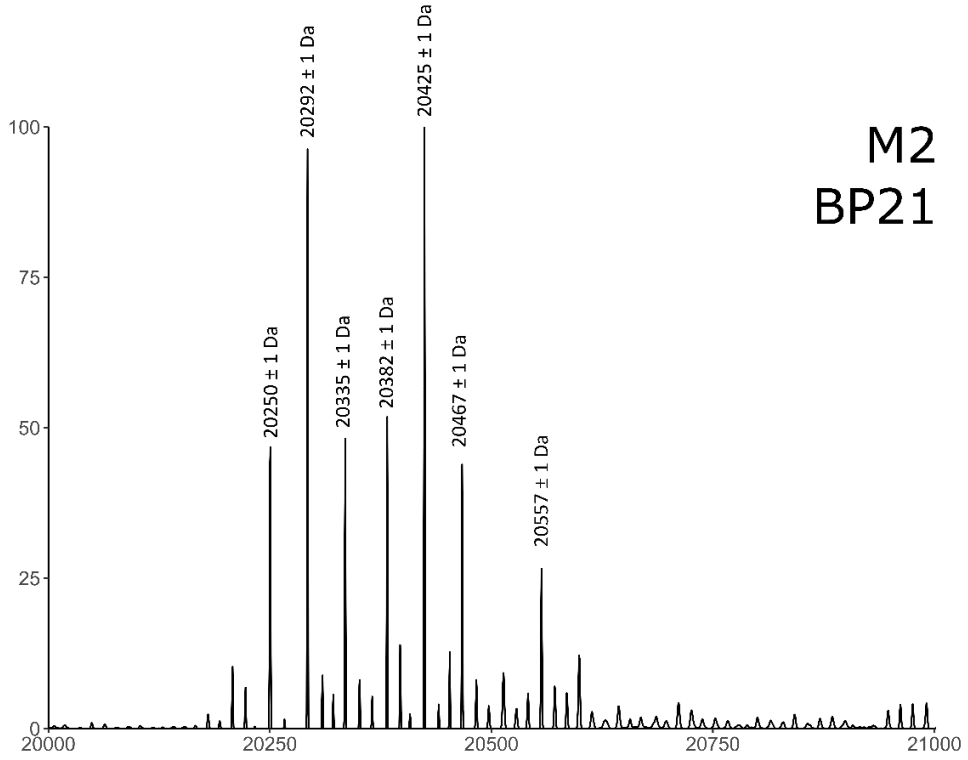


Figure 5.3 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 2 (male 2, sample BP21).

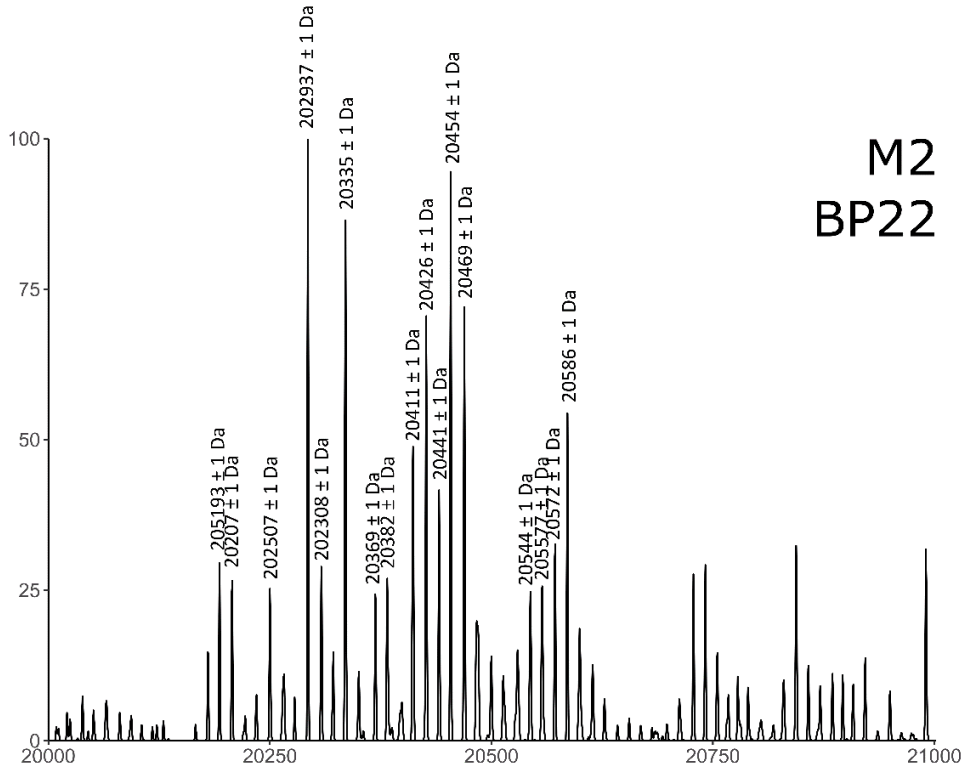


Figure 5.4 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 2 (male 2, sample BP22).

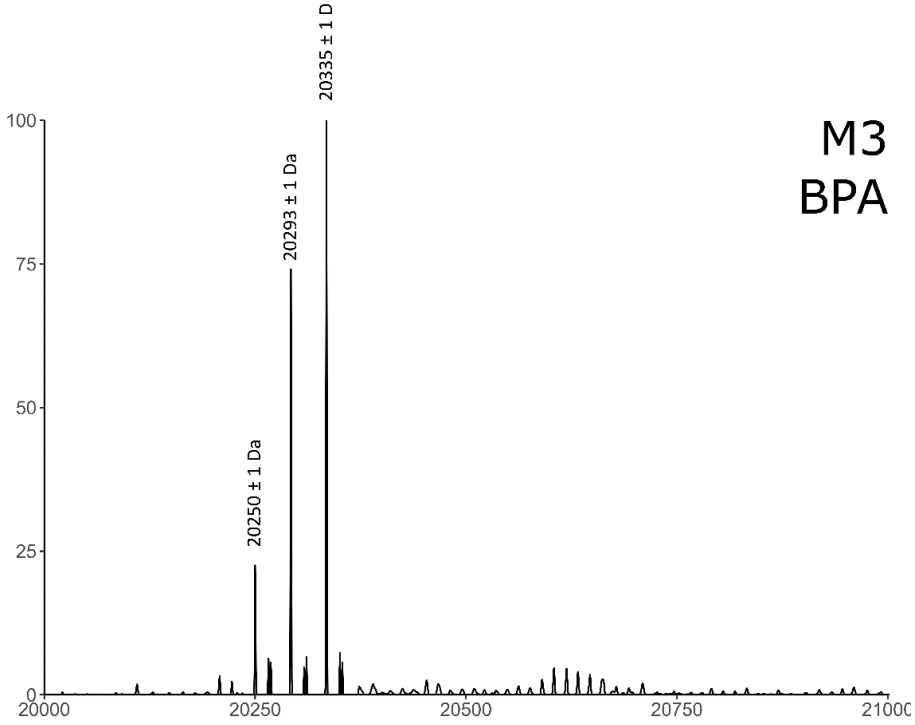


Figure 5.5 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 3 (male 3, sample BPA).

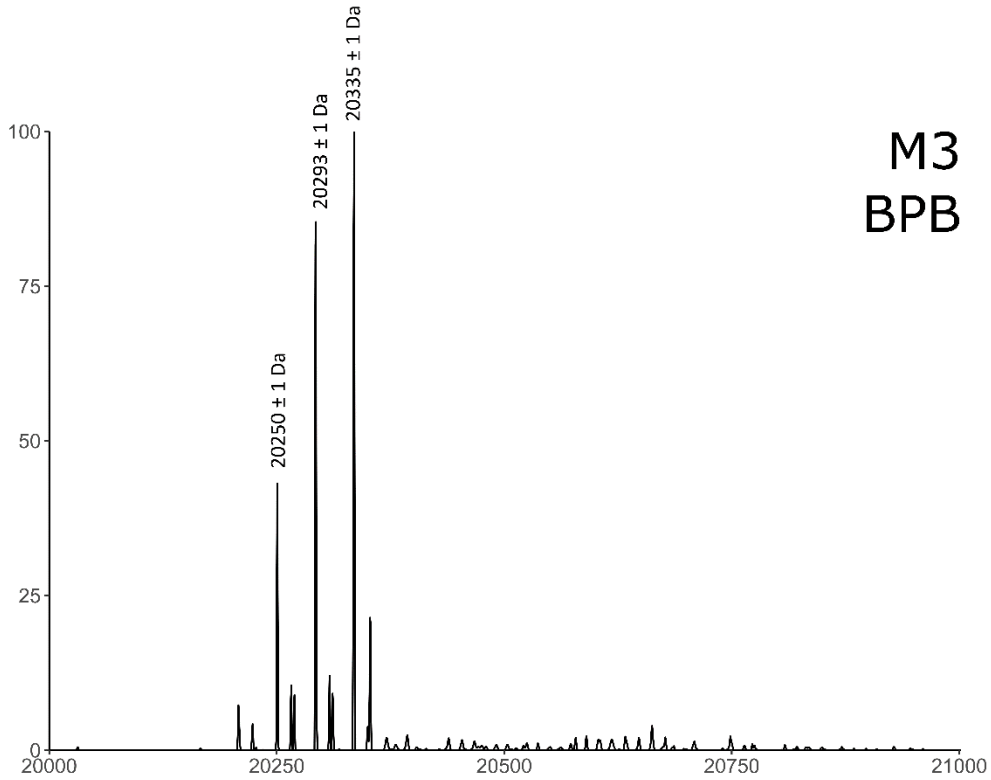


Figure 5.6 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 3 (male 3, sample BPB).

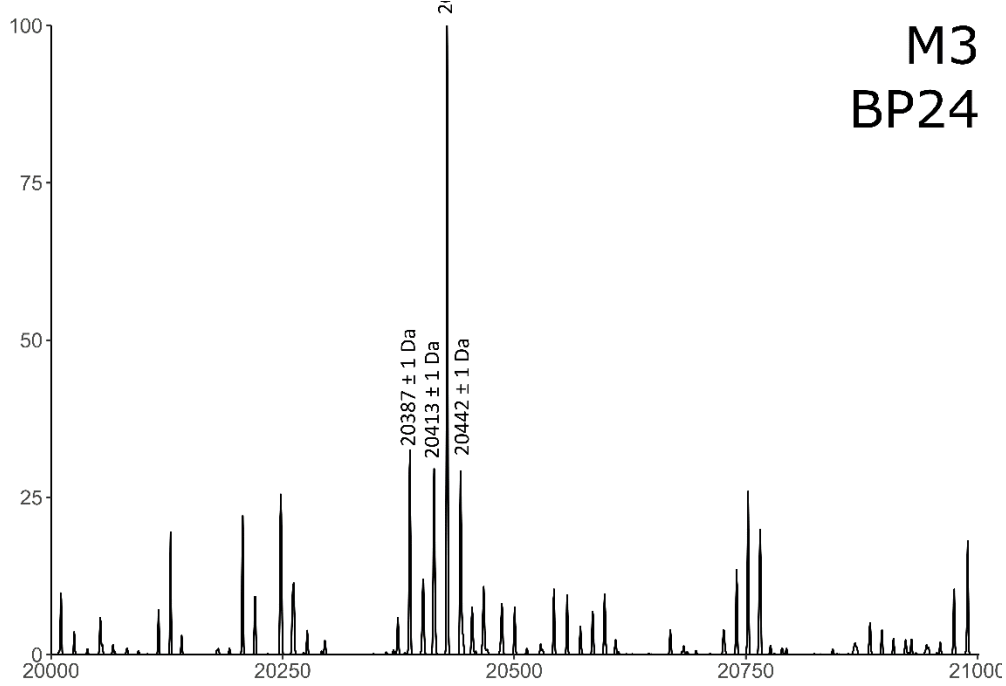


Figure 5.7 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 3 (male 3, sample BP24).

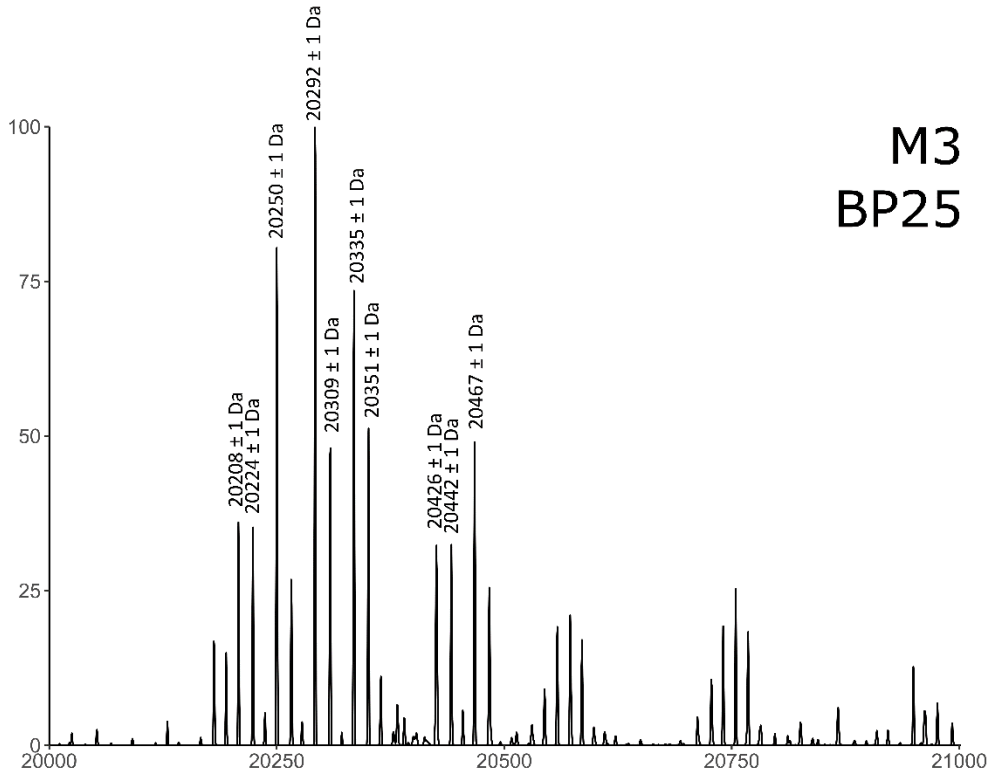


Figure 5.8 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 3 (male 3, sample BP25).

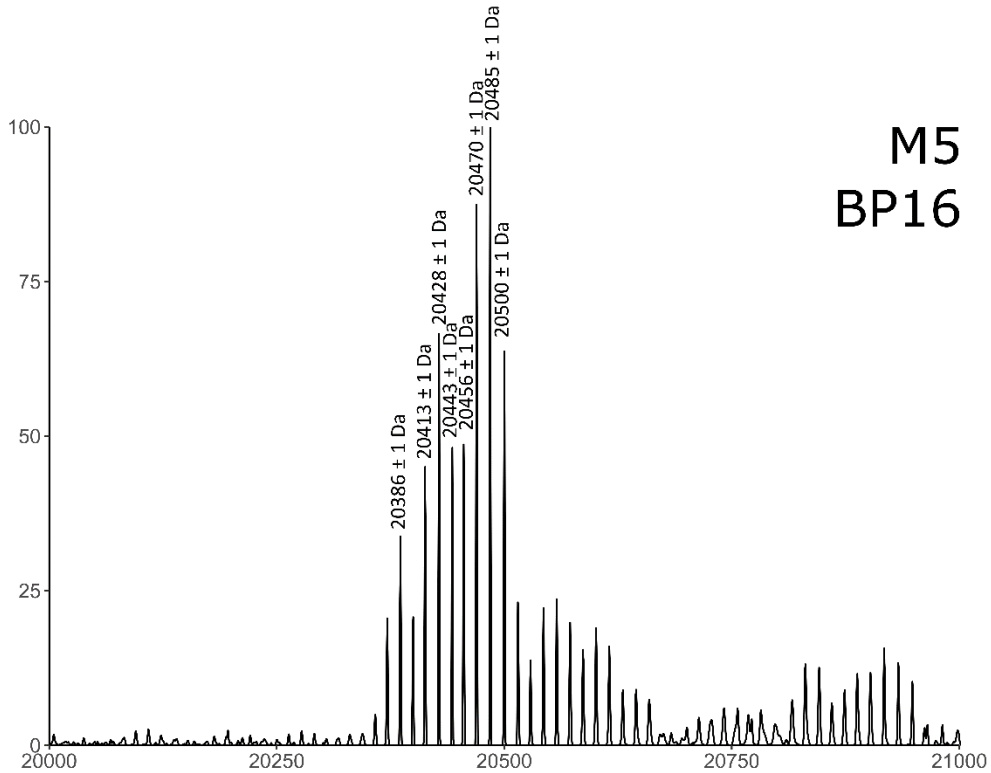


Figure 5.9 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 5 (male 5, sample BP16).

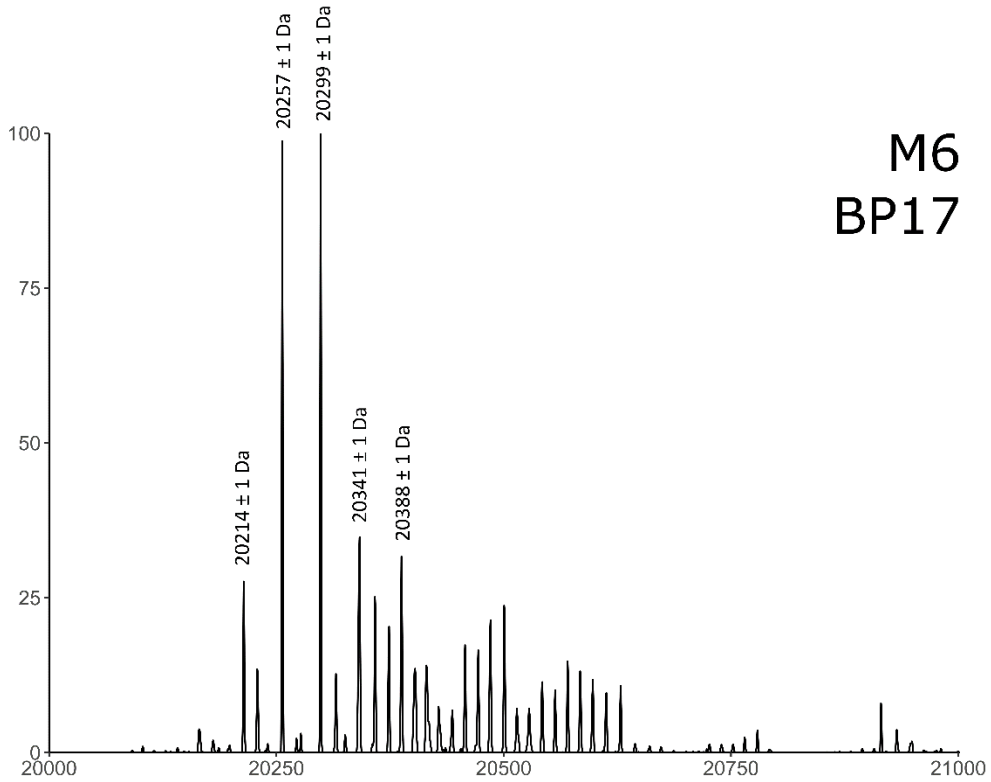


Figure 5.10 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of male brushtail possum 6 (male 6, sample BP17).

FEMALES

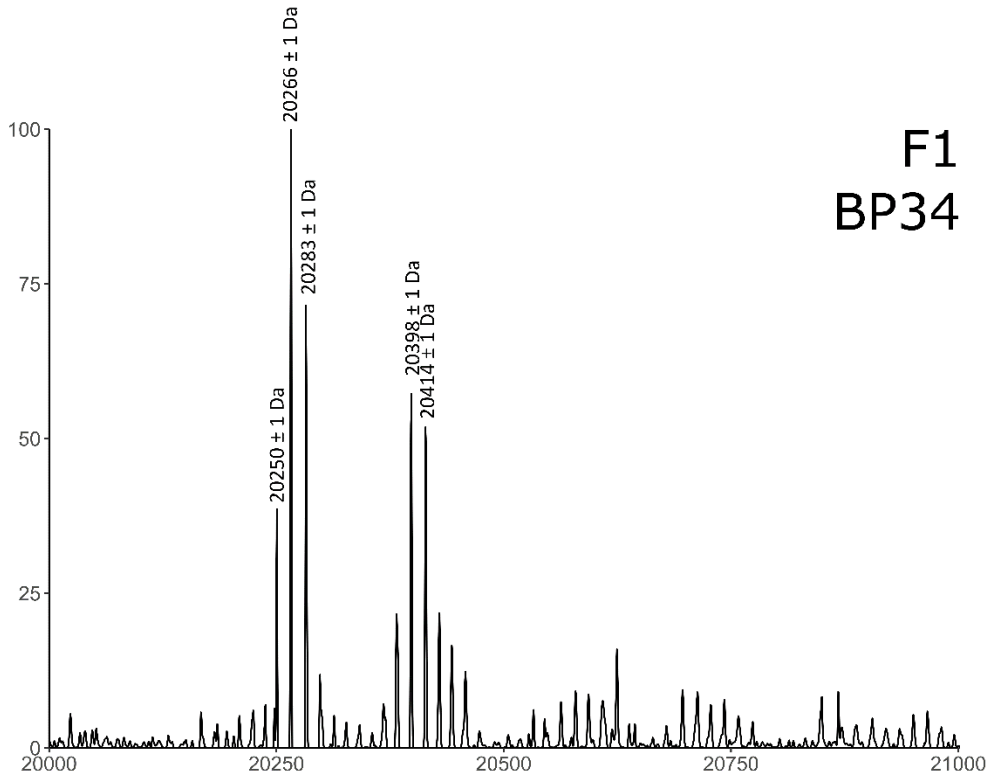


Figure 5.11 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 1 (female 1, sample BP34).

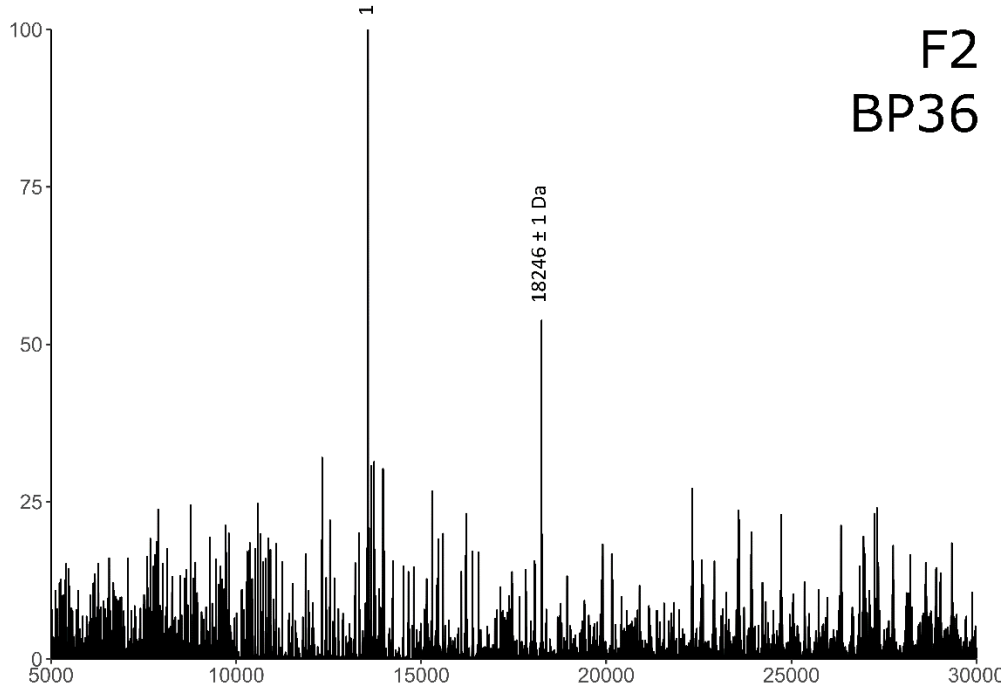


Figure 5.12 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 2 (female 2, sample BP36).

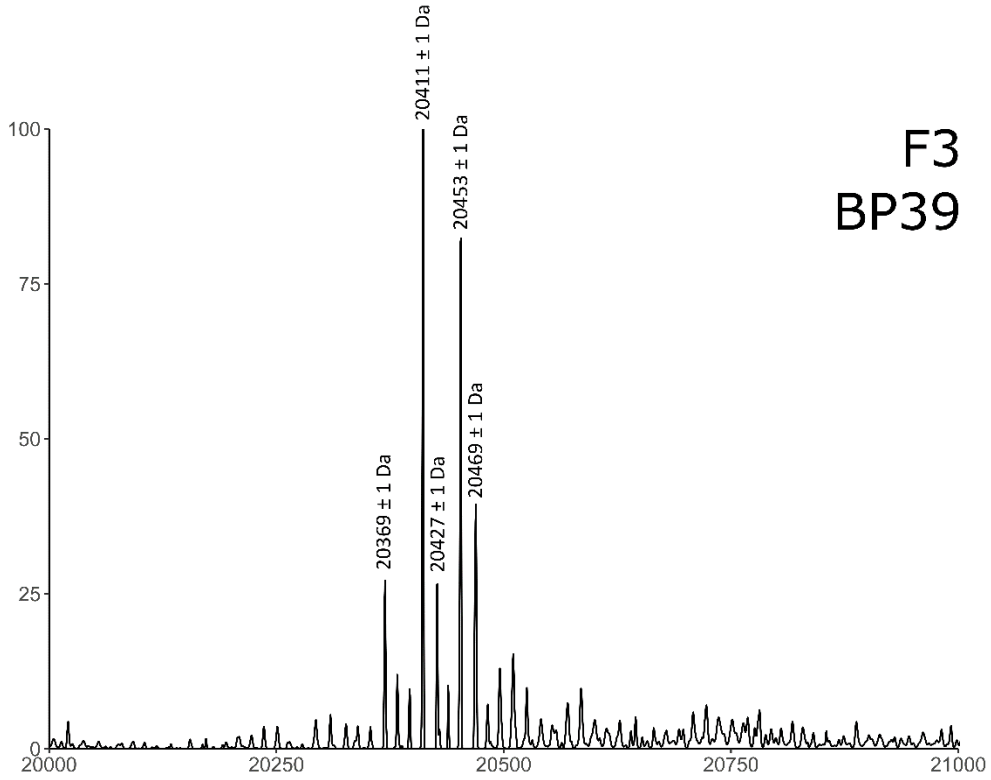


Figure 5.13 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 3 (female 3, sample BP39).

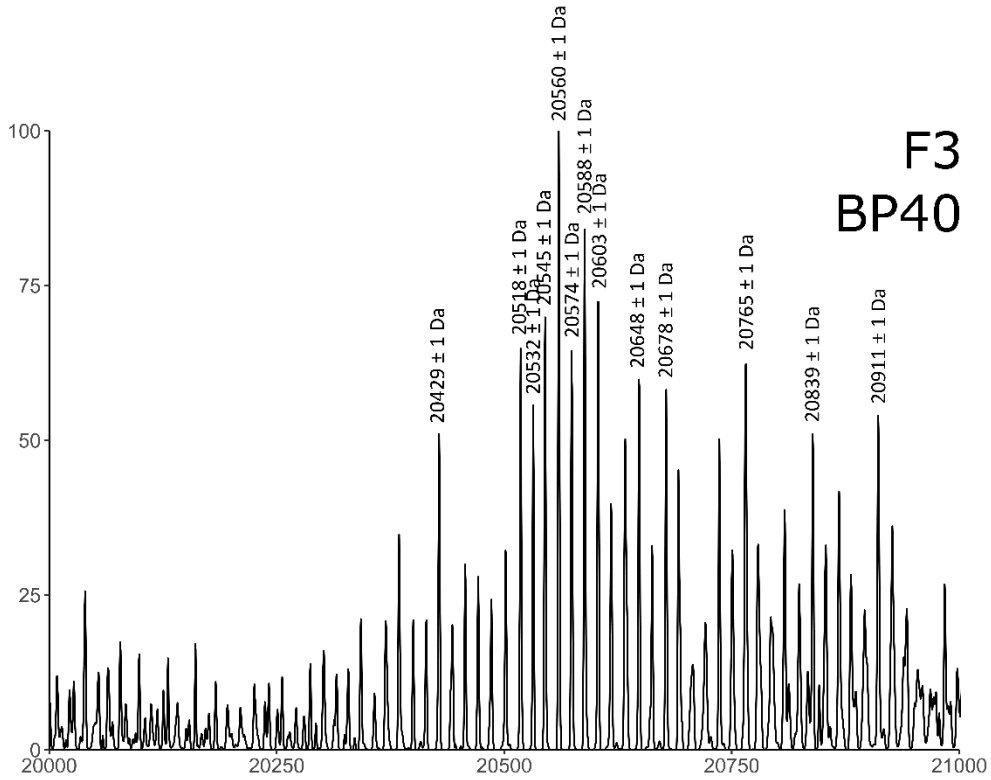


Figure 5.14 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 3 (female 1, sample BP40).

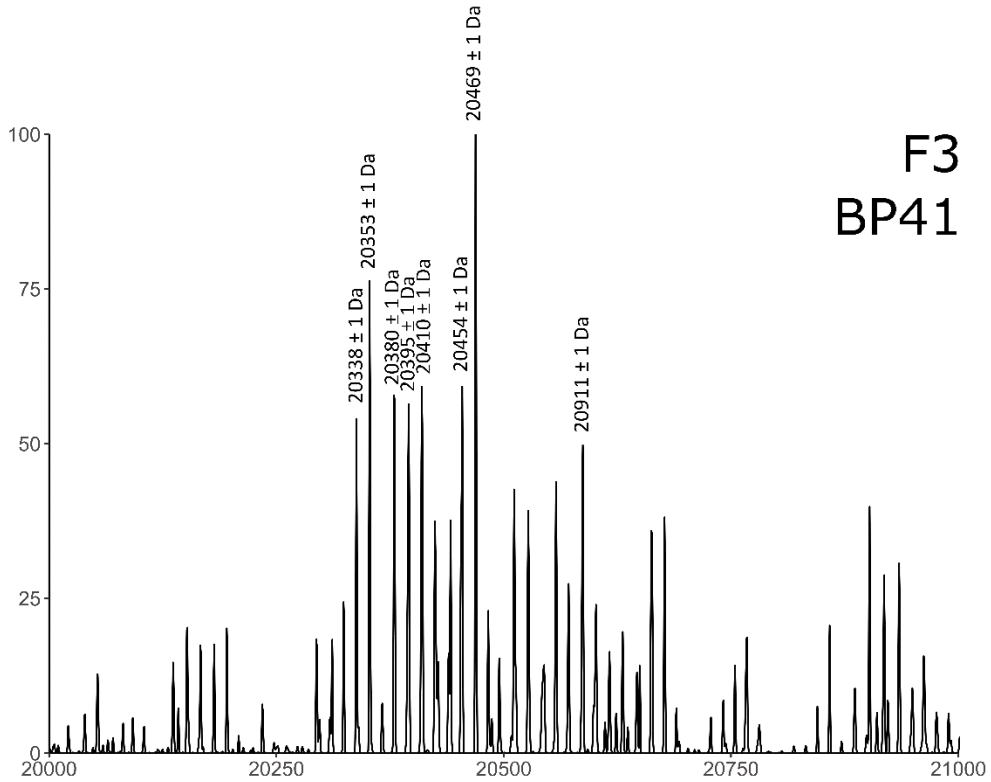


Figure 5.15 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 3 (female 3, sample BP41).

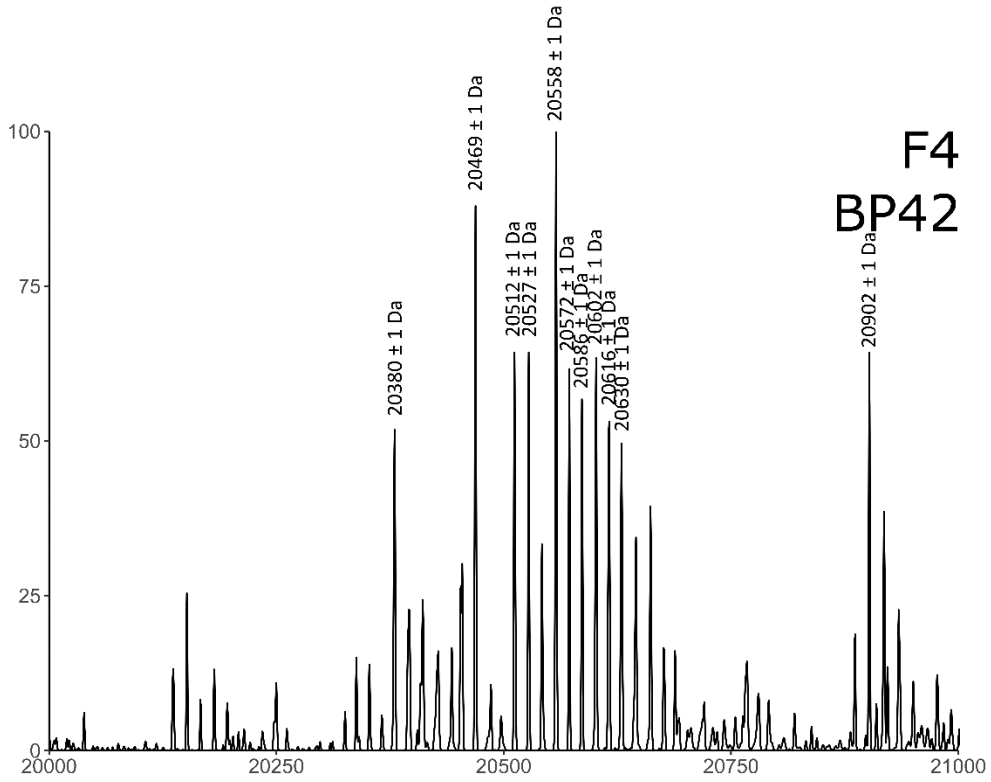


Figure 5.16 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 4 (female 4, sample BP42).

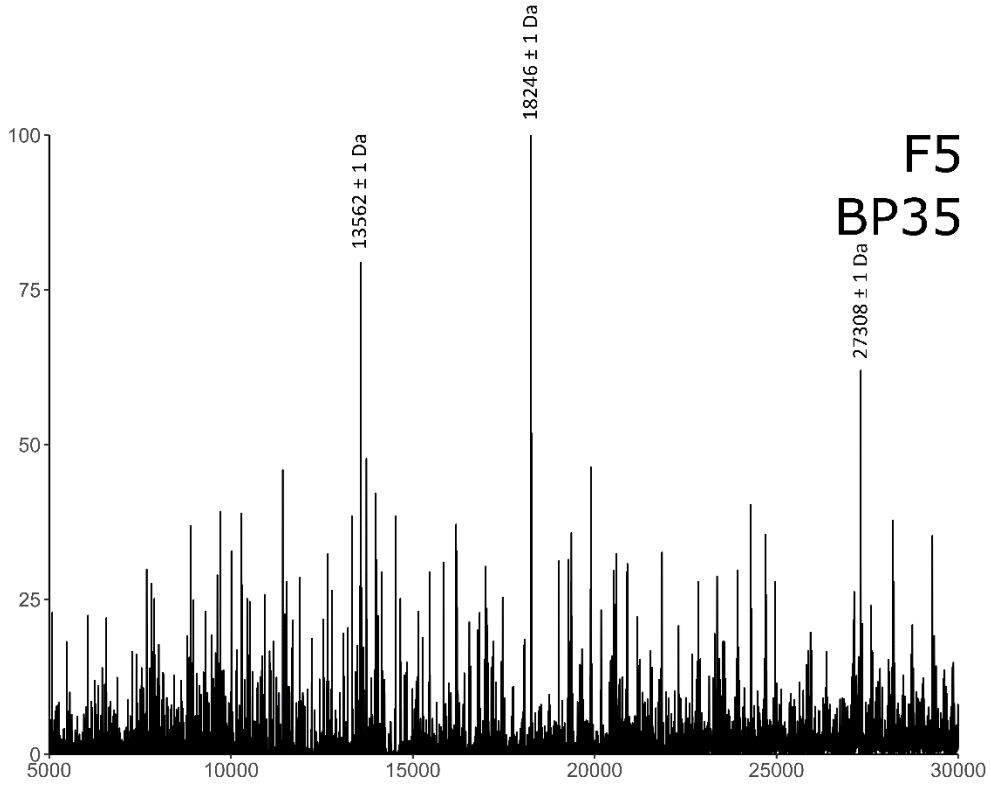


Figure 5.17 Deconvoluted mass spectrum generated by analysis of intact proteins in the urine of female brushtail possum 5 (female 5, sample BP35).

5.6.2 *MALDI-ToF peptide mass fingerprints from tryptic in-gel digestion*

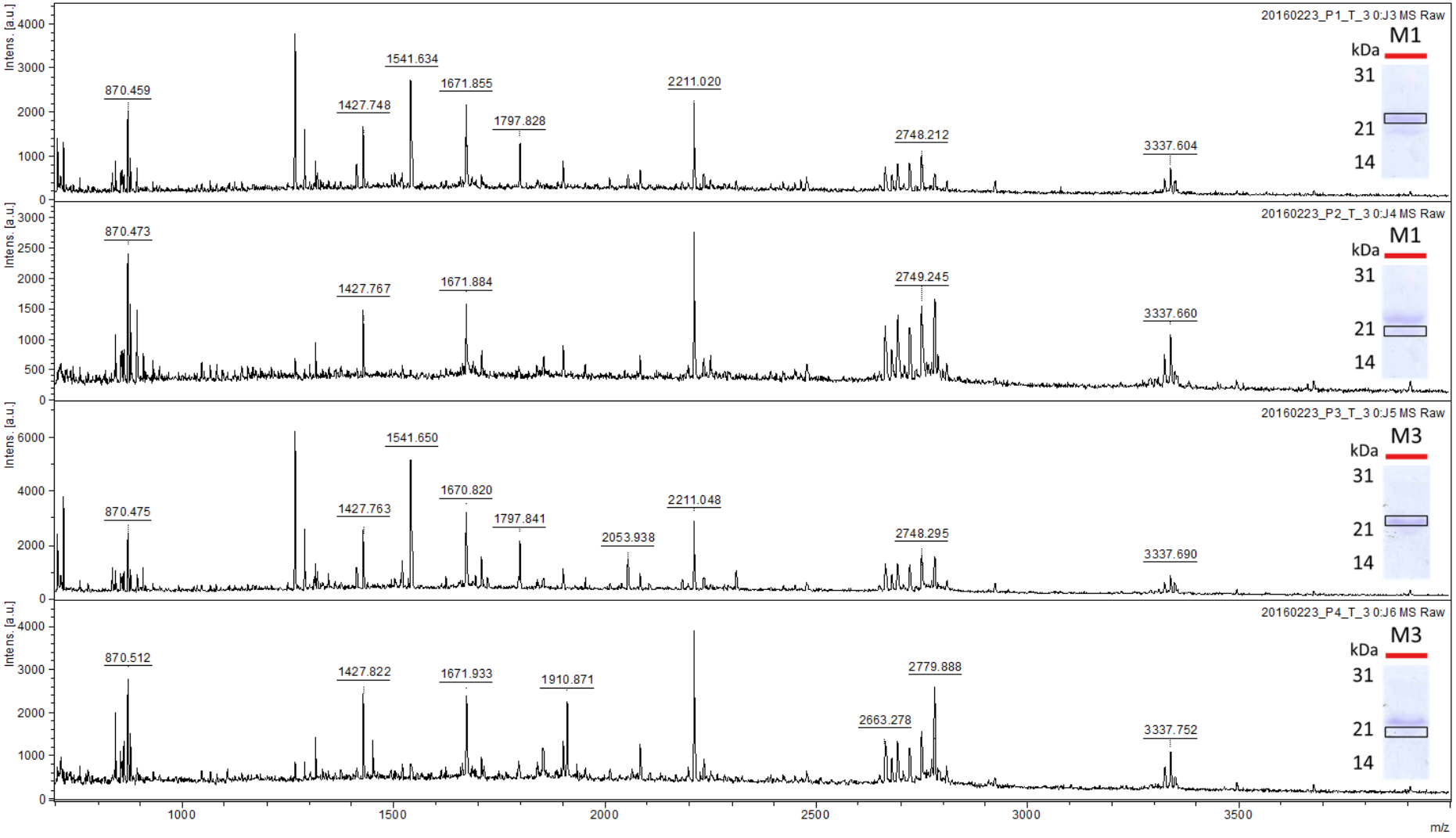


Figure 5.18 Mass spectra generated by MALDI-ToF analysis of urinary proteins separated by SDS-PAGE and digested in-gel. Protein bands resolving at approximately 25 kDa and 20 kDa were analysed, from Male 1 and Male 3.

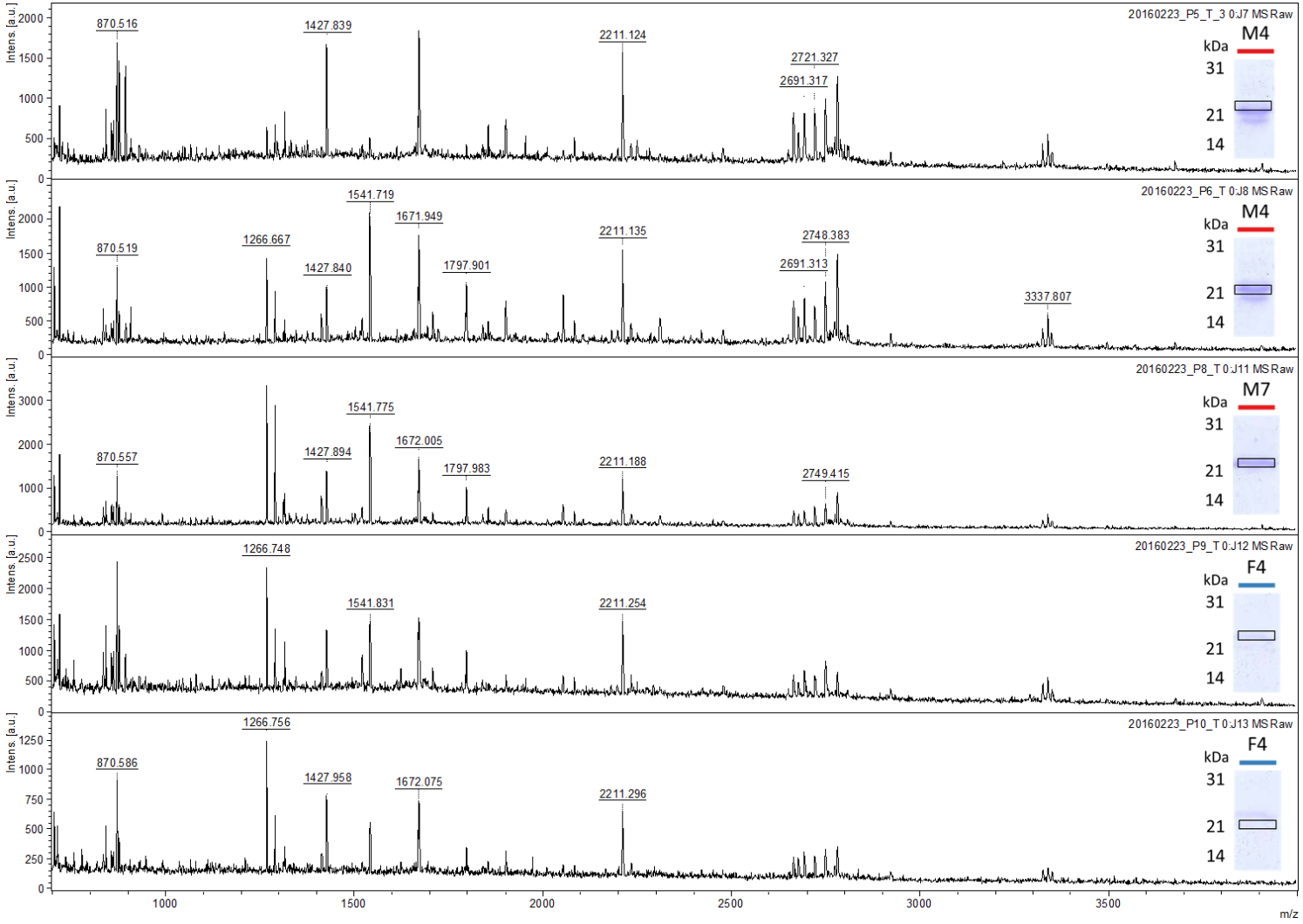


Figure 5.19 Mass spectra generated by MALDI-ToF analysis of urinary proteins separated by SDS-PAGE and digested in-gel. Protein bands resolving at approximately 25 kDa and 20 kDa were analysed, from Male 4, Male 7 and Female 4.

5.6.3 *Homologous sequences to the trichosurin-like protein from* M. domestica

Table 5.1 Sequences identified from searching the NCBI and UniProt databases as homologous to the trichosurin-like protein from *M. domestica* that were used as a guide to sequence vulpeculin *de novo*.

ACCESSION	FULL DESCRIPTION	SPECIES	COMMON NAME
2R73_A	Chain A, Crystal Structure Of The Possum Milk Whey Lipocalin Trichosurin At Ph 8.2	*Trichosurus vulpecula*	Common Brushtail Possum
XP_012396091.1	trichosurin-like	*Sarcophilus harrisii*	Tasmanian Devil
XP_023351137.1	minor allergen Can f 2-like	*Sarcophilus harrisii*	Tasmanian Devil
XP_003758835.1	trichosurin-like isoform X1	*Sarcophilus harrisii*	Tasmanian Devil
XP_012396094.1	trichosurin-like isoform X2	*Sarcophilus harrisii*	Tasmanian Devil
XP_007475411.1	PREDICTED: trichosurin-like isoform X1	*Monodelphis domestica*	Grey short-tailed opossum
XP_007475412.1	PREDICTED: trichosurin-like isoform X2	*Monodelphis domestica*	Grey short-tailed opossum
XP_007475413.1	PREDICTED: trichosurin-like	*Monodelphis domestica*	Grey short-tailed opossum
XP_020837036.1	major urinary protein-like	*Phascolarctos cinereus*	Koala
XP_020837035.1	major urinary protein-like isoform X2	*Phascolarctos cinereus*	Koala
XP_020837033.1	trichosurin-like isoform X1	*Phascolarctos cinereus*	Koala
XP_020837067.1	epididymal-specific lipocalin-9	*Phascolarctos cinereus*	Koala
XP_019491344.1	PREDICTED: allergen Fel d 4-like	*Hipposideros armiger*	Great Roundleaf Bat
XP_012878672.1	PREDICTED: allergen Fel d 4-like	*Dipodomys ordii*	Ord's Kangaroo Rat
XP_004678007.2	PREDICTED: major urinary protein 4-like	*Condylura cristata*	Star-Nosed Mole
AAA88508.1	alpha-2u globulin	*Rattus norvegicus*	Brown Rat
3ZQ3_A	Chain A, Crystal Structure Of Rat Odorant Binding Protein 3 (obp3)	*Rattus norvegicus*	Brown Rat
XP_008761838.1	PREDICTED: major urinary protein-like	*Rattus norvegicus*	Brown Rat
XP_006238283.2	PREDICTED: major urinary protein isoform X1	*Rattus norvegicus*	Brown Rat
XP_024417856.1	allergen Fel d 4-like isoform X1	*Desmodus rotundus*	Common Vampire Bat
XP_024417857.1	allergen Fel d 4-like isoform X2	*Desmodus rotundus*	Common Vampire Bat
XP_008587887.1	PREDICTED: major allergen Equ c 1-like	*Galeopterus variegatus*	Sunda Flying Lemur
XP_004747887.1	PREDICTED: epididymal-specific lipocalin-9	*Mustela putorius furo*	Ferret
XP_005329807.1	major urinary protein 20-like	*Ictidomys tridecemlineatus*	Thirteen-Lined Ground Squirrel
XP_015347583.1	PREDICTED: major urinary protein 20-like	*Marmota marmota marmota*	Alpine Marmot

5.6.4 *Fragment ion spectra for sequencing*

The following fragment ion spectra are labelled according to the peptide coverage map below of the first draft sequence of vulpeculin, by peptide order from N-terminus to C-terminus, according to protease (i.e. the glu-C peptide closest to the N-terminus is denoted g1).

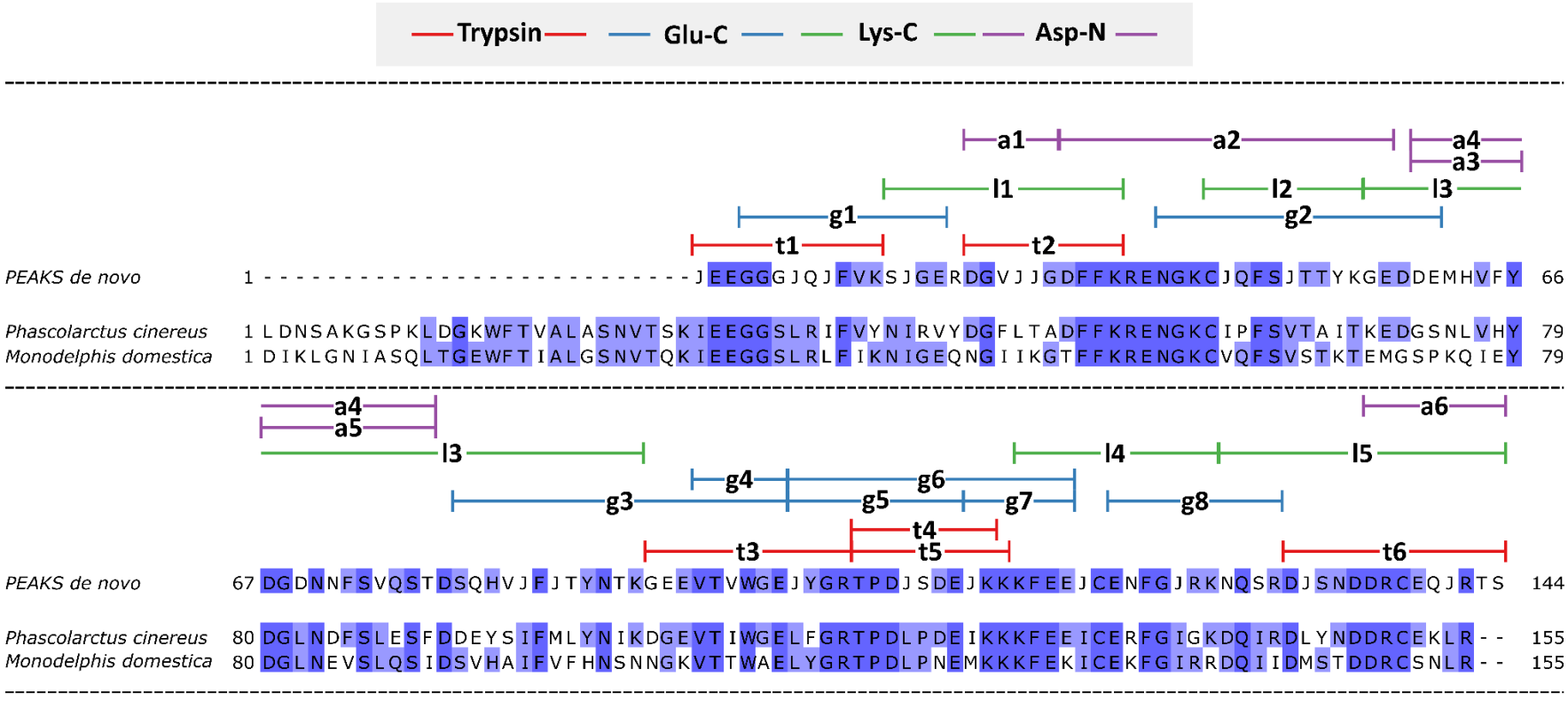
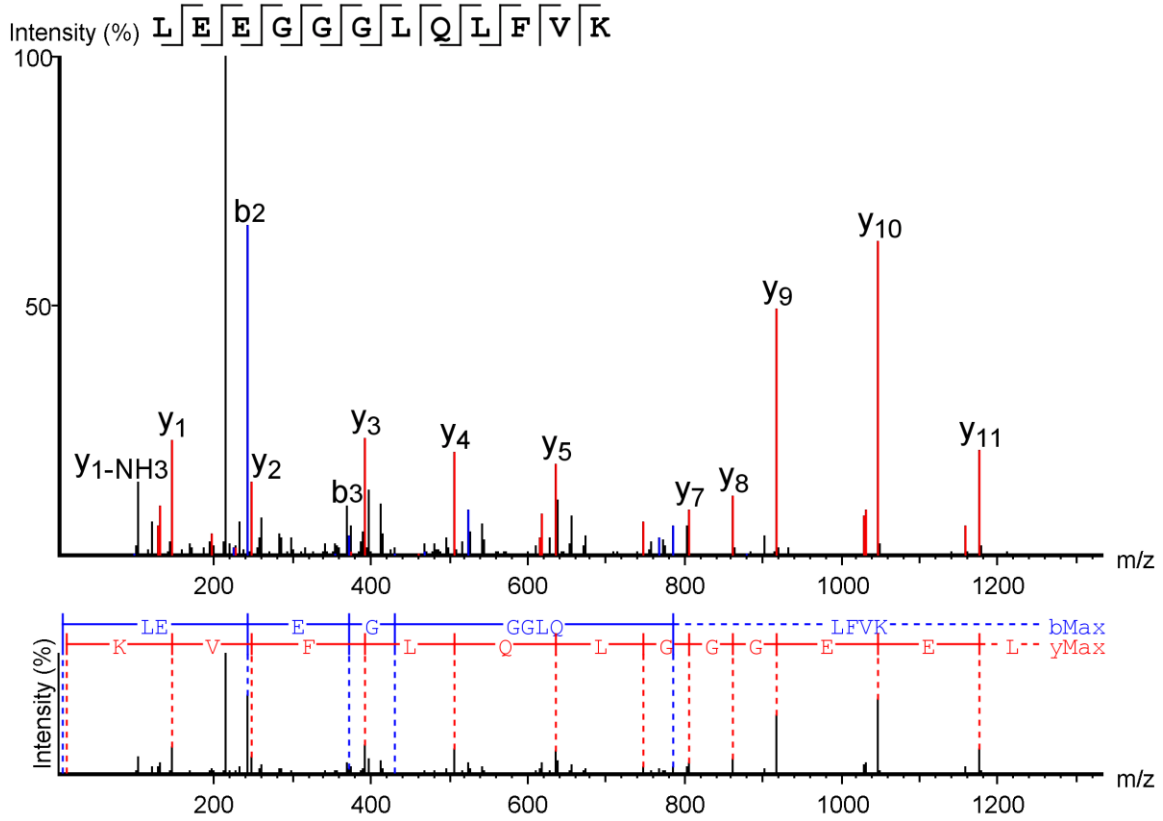
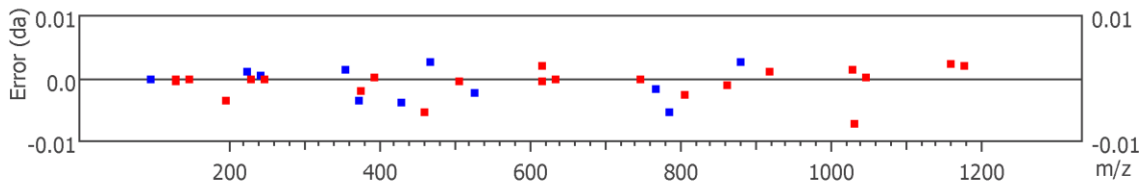


Figure 5.20 Peptide coverage map of vulpeculin (denoted 'PEAKS de novo'), from peptides generated with trypsin (t1-6), glu-C (g1-8), lys-C (l1-5) and asp-N (a1-6)., aligned with trichosurin-like protein (*Monodelphis domestica*, short-tailed oppossum) and mup-like protein (*Phascolarctus cinereus*, koala) identified as homologous proteins from a search of all mammalian protein sequences in the NCBI database using BLAST, and aligned in JalView [https://www.jalview.org/].

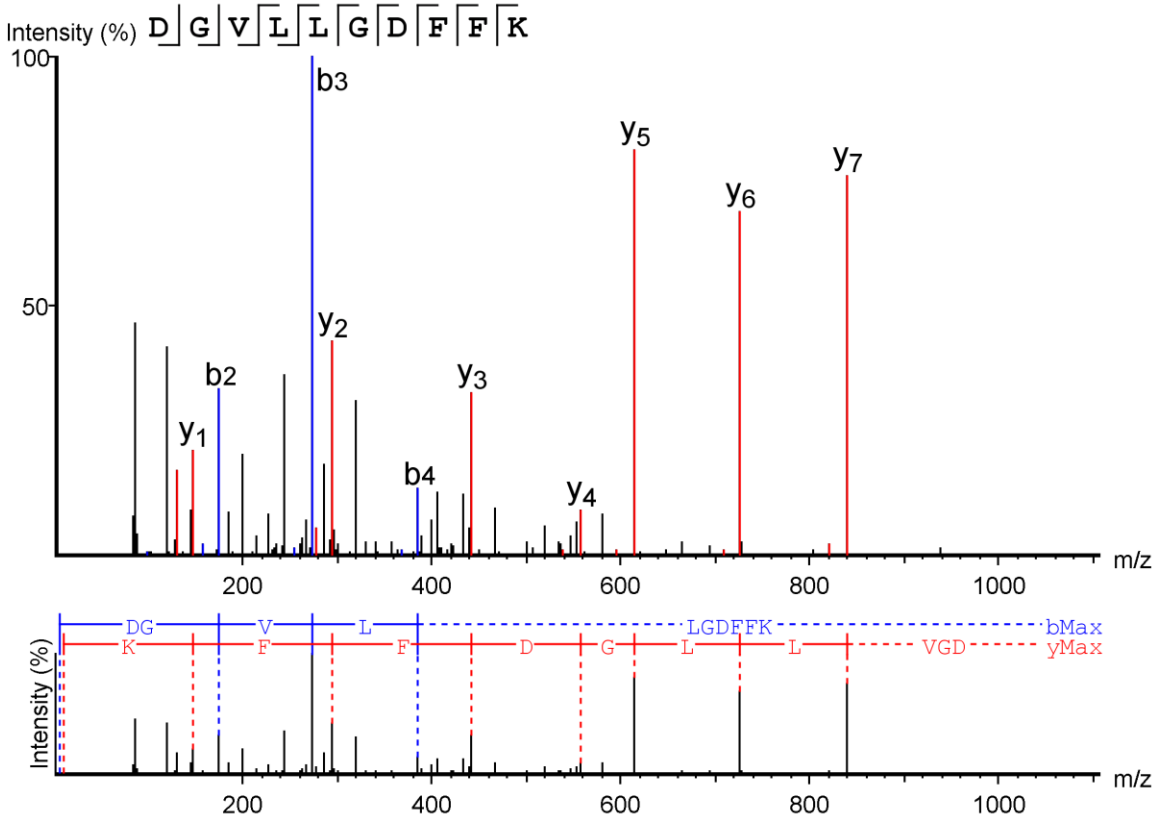
t1



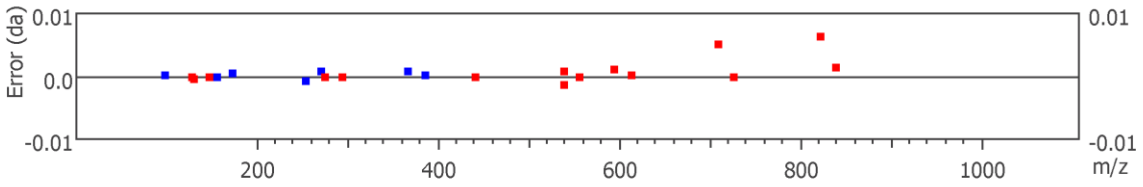
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0648	57.5459	L					12
2	243.1337	225.1226	226.1075	122.0672	E	1176.6234	1158.6127	1159.5989	588.8129	11
3	372.1807	354.1648	355.1501	186.5885	E	1047.5829	1029.5710	1030.5636	524.2916	10
4	429.2023	411.1880	412.1715	215.0993	G	918.5392	900.5301	901.5137	459.7758	9
5	486.2200	468.2066	469.1930	243.6100	G	861.5202	843.5087	844.4922	431.2596	8
6	543.2415	525.2331	526.2145	272.1207	G	804.5002	786.4872	787.4708	402.7489	7
7	656.3255	638.3149	639.2985	328.6628	L	747.4763	729.4658	730.4493	374.2401	6
8	784.3893	766.3751	767.3589	392.6920	Q	634.3923	616.3793	617.3658	317.6961	5
9	897.4681	879.4547	880.4412	449.2341	L	506.3340	488.3231	489.3067	253.6668	4
10	1044.5365	1026.5260	1027.5095	522.7682	F	393.2492	375.2390	376.2226	197.1283	3
11	1143.6050	1125.5945	1126.5780	572.3025	V	246.1810	228.1709	229.1542	123.5906	2
12					K	147.1128	129.1022	130.0862	74.0564	1



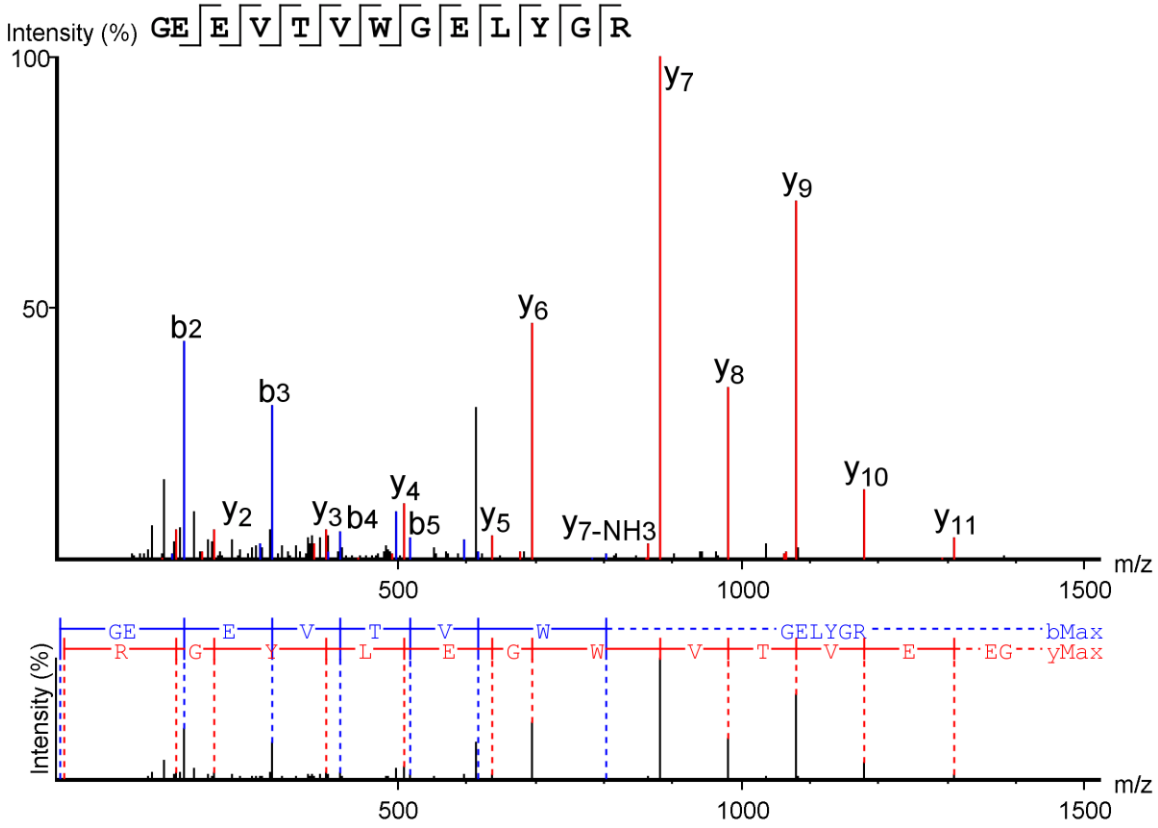
t2



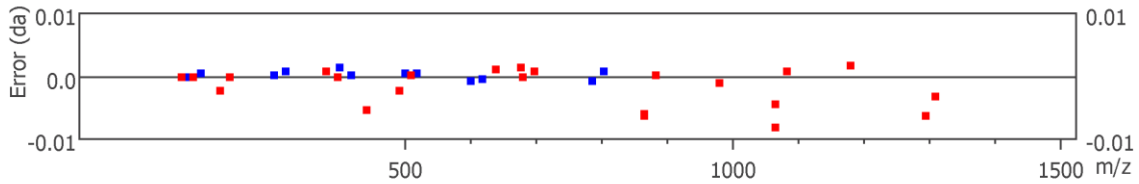
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0238	99.0078	58.5174	D					10
2	173.0554	155.0457	156.0291	87.0281	G	995.5560	977.5454	978.5290	498.2780	9
3	272.1236	254.1149	255.0977	136.5623	V	938.5345	920.5240	921.5076	469.7673	8
4	385.2082	367.1971	368.1817	193.1043	L	839.4643	821.4489	822.4391	420.2331	7
5	498.2928	480.2822	481.2658	249.6464	L	726.3818	708.3661	709.3551	363.6910	6
6	555.3142	537.3036	538.2872	278.1571	G	613.2974	595.2860	596.2710	307.1490	5
7	670.3411	652.3306	653.3141	335.6706	D	556.2769	538.2648	539.2511	278.6383	4
8	817.4095	799.3990	800.3826	409.2048	F	441.2495	423.2390	424.2226	221.1248	3
9	964.4780	946.4674	947.4510	482.7390	F	294.1810	276.1709	277.1542	147.5906	2
10					K	147.1126	129.1021	130.0864	74.0564	1



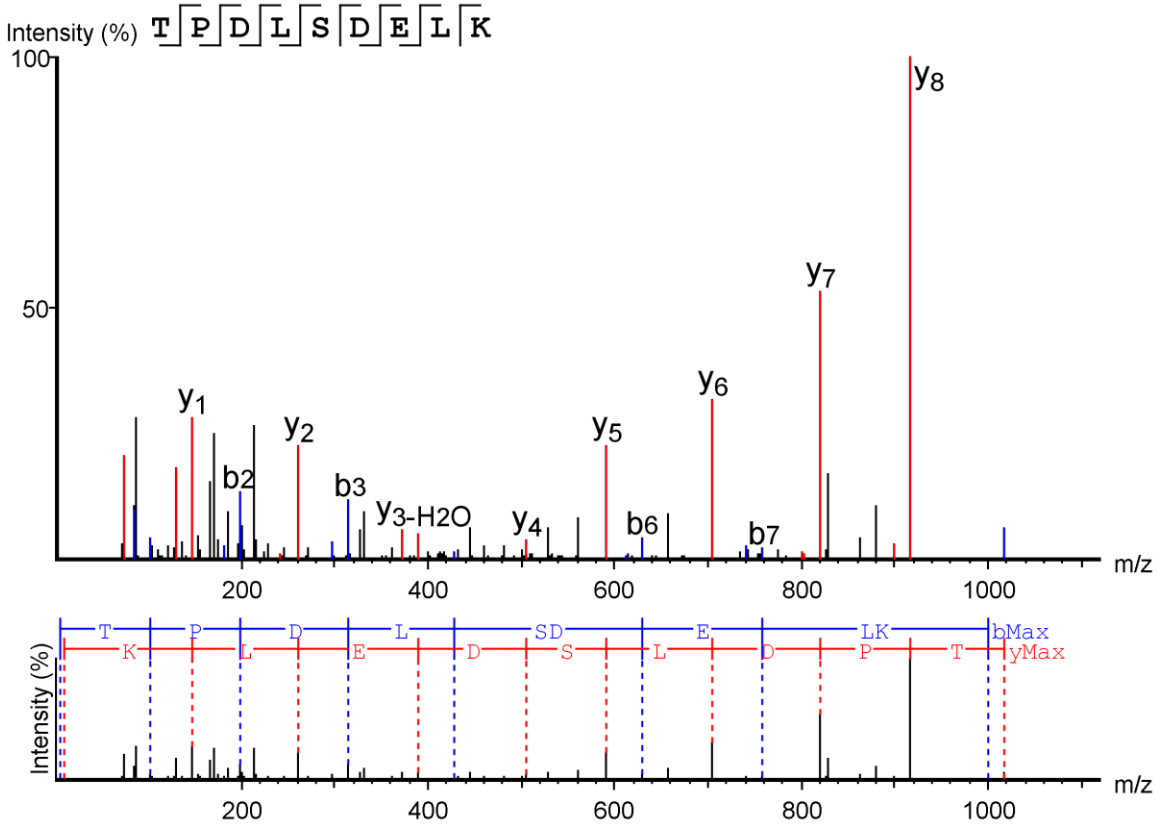
t3



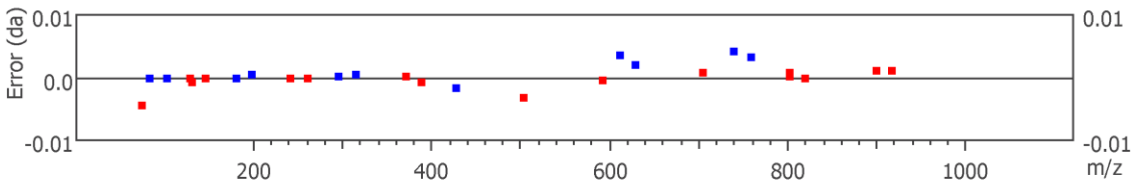
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					13
2	187.0713	169.0613	170.0450	94.0359	E	1437.7008	1419.6903	1420.6738	719.3504	12
3	316.1135	298.1033	299.0875	158.5572	E	1308.6614	1290.6477	1291.6375	654.8291	11
4	415.1824	397.1705	398.1559	208.0914	V	1179.6136	1161.6051	1162.5886	590.3078	10
5	516.2297	498.2193	499.2036	258.6153	T	1080.5461	1062.5411	1063.5286	540.7736	9
6	615.2994	597.2891	598.2720	308.1495	V	979.5007	961.4890	962.4726	490.2498	8
7	801.3771	783.3687	784.3513	401.1891	W	880.4308	862.4268	863.4101	440.7209	7
8	858.3997	840.3892	841.3727	429.6999	G	694.3506	676.3395	677.3251	347.6759	6
9	987.4423	969.4318	970.4153	494.2212	E	637.3289	619.3198	620.3034	319.1652	5
10	1100.5264	1082.5159	1083.4994	550.7632	L	508.2874	490.2772	491.2631	254.6439	4
11	1263.5897	1245.5792	1246.5627	632.2949	Y	395.2038	377.1931	378.1757	198.1019	3
12	1320.6112	1302.6007	1303.5842	660.8056	G	232.1405	214.1298	215.1156	116.5702	2
13					R	175.1192	157.1087	158.0919	88.0595	1



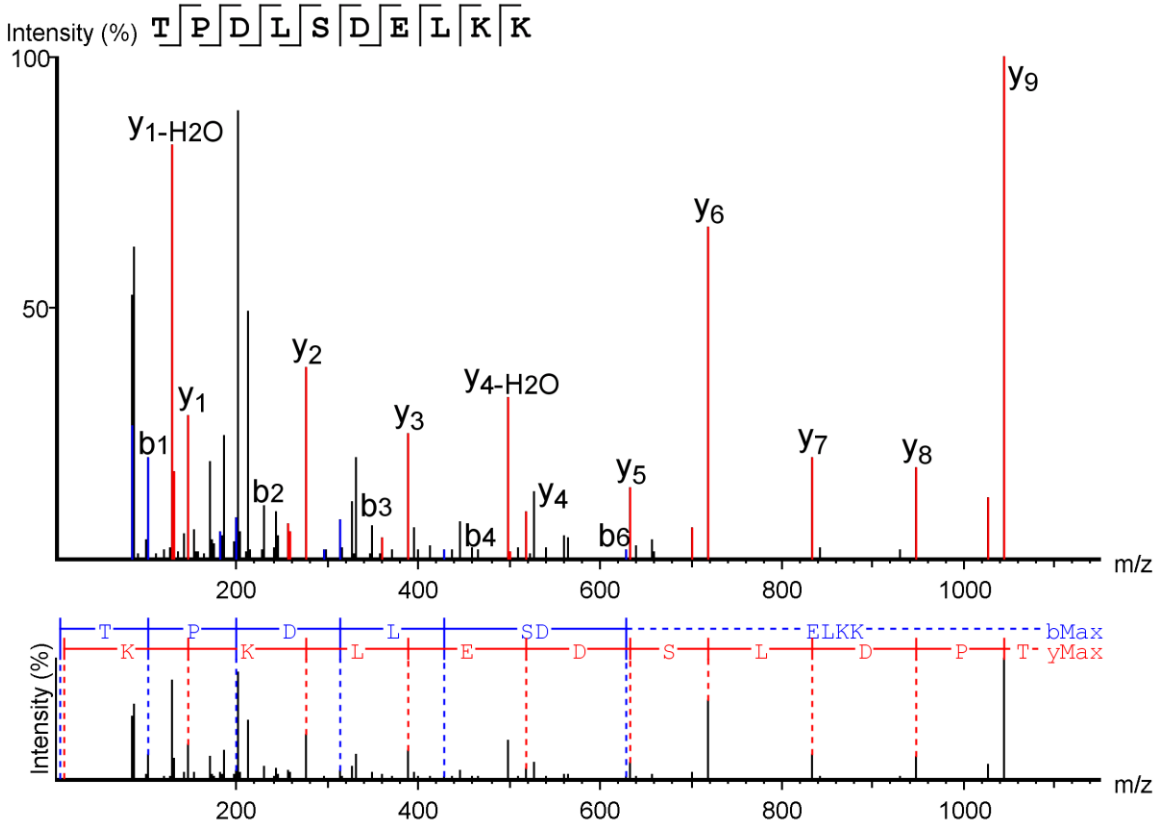
t4



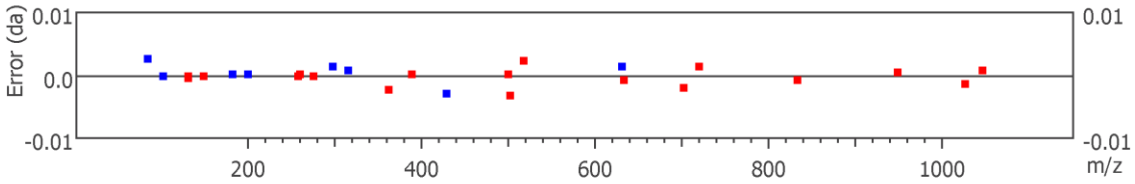
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	102.0555	84.0450	85.0285	51.5278	T					9
2	199.1075	181.0976	182.0813	100.0541	P	916.4608	898.4502	899.4352	458.7311	8
3	314.1346	296.1242	297.1082	157.5676	D	819.4092	801.3978	802.3821	410.2047	7
4	427.2211	409.2087	410.1923	214.1096	L	704.3815	686.3719	687.3555	352.6912	6
5	514.2513	496.2407	497.2243	257.6256	S	591.2989	573.2878	574.2714	296.1492	5
6	629.2760	611.2639	612.2512	315.1391	D	504.2695	486.2558	487.2394	252.6332	4
7	758.3173	740.3058	741.2938	379.6604	E	389.2403	371.2284	372.2125	195.1197	3
8	871.4048	853.3943	854.3779	436.2024	L	260.1967	242.1862	243.1698	130.5984	2
9					K	147.1129	129.1021	130.0865	74.0607	1



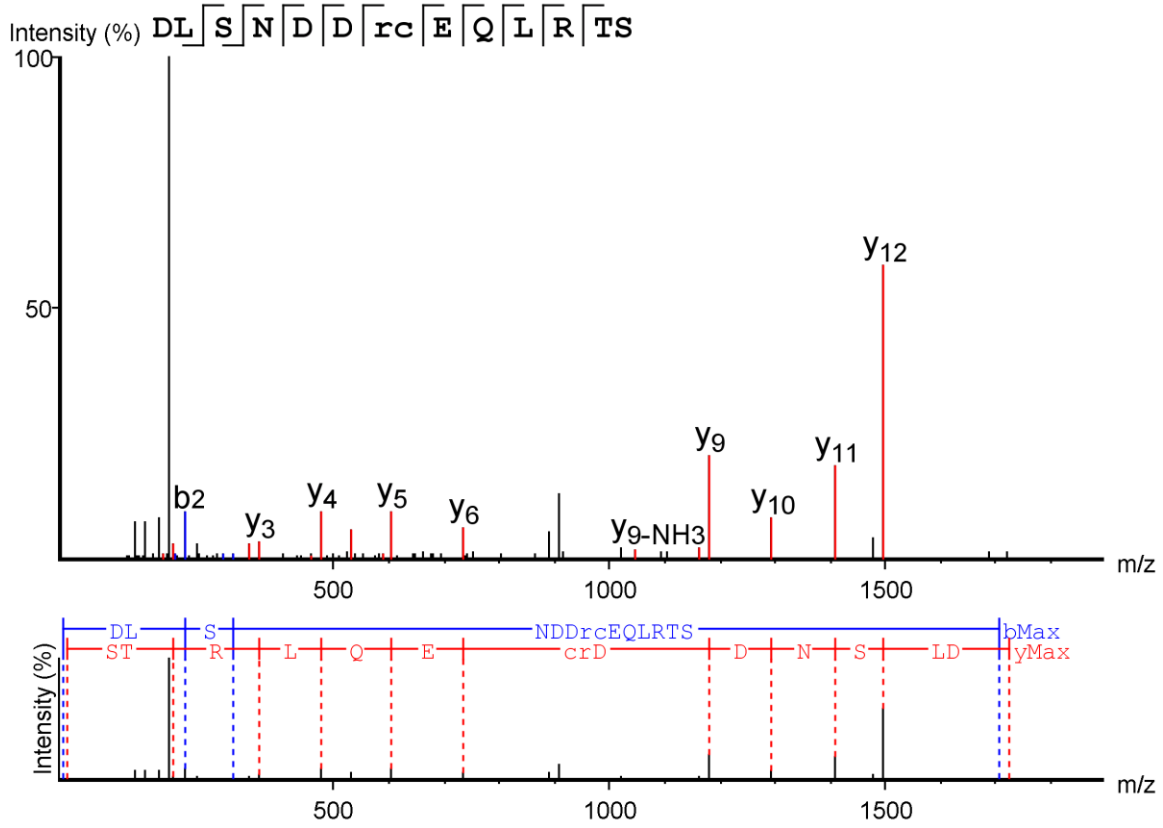
t5



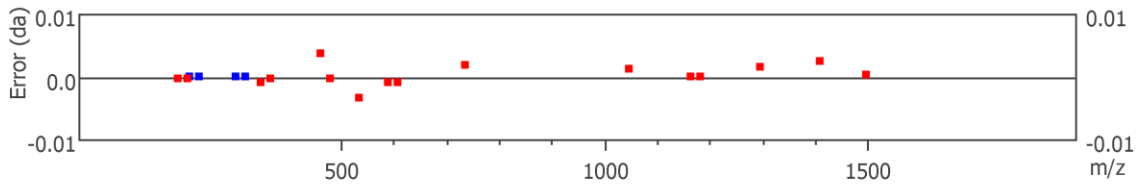
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	102.0554	84.0419	85.0285	51.5278	T					10
2	199.1078	181.0971	182.0813	100.0541	P	1044.5562	1026.5481	1027.5302	522.7786	9
3	314.1340	296.1230	297.1082	157.5676	D	947.5035	929.4938	930.4774	474.2522	8
4	427.2222	409.2087	410.1923	214.1096	L	832.4781	814.4669	815.4504	416.7387	7
5	514.2513	496.2407	497.2243	257.6256	S	719.3917	701.3848	702.3664	360.1989	6
6	629.2766	611.2676	612.2512	315.1391	D	632.3621	614.3508	615.3344	316.6807	5
7	758.3208	740.3102	741.2938	379.6604	E	517.3318	499.3234	500.3106	259.1672	4
8	871.4048	853.3943	854.3779	436.2024	L	388.2914	370.2812	371.2648	194.6459	3
9	999.4998	981.4893	982.4728	500.2499	K	275.2079	257.1971	258.1803	138.1039	2
10					K	147.1129	129.1021	130.0863	74.0564	1



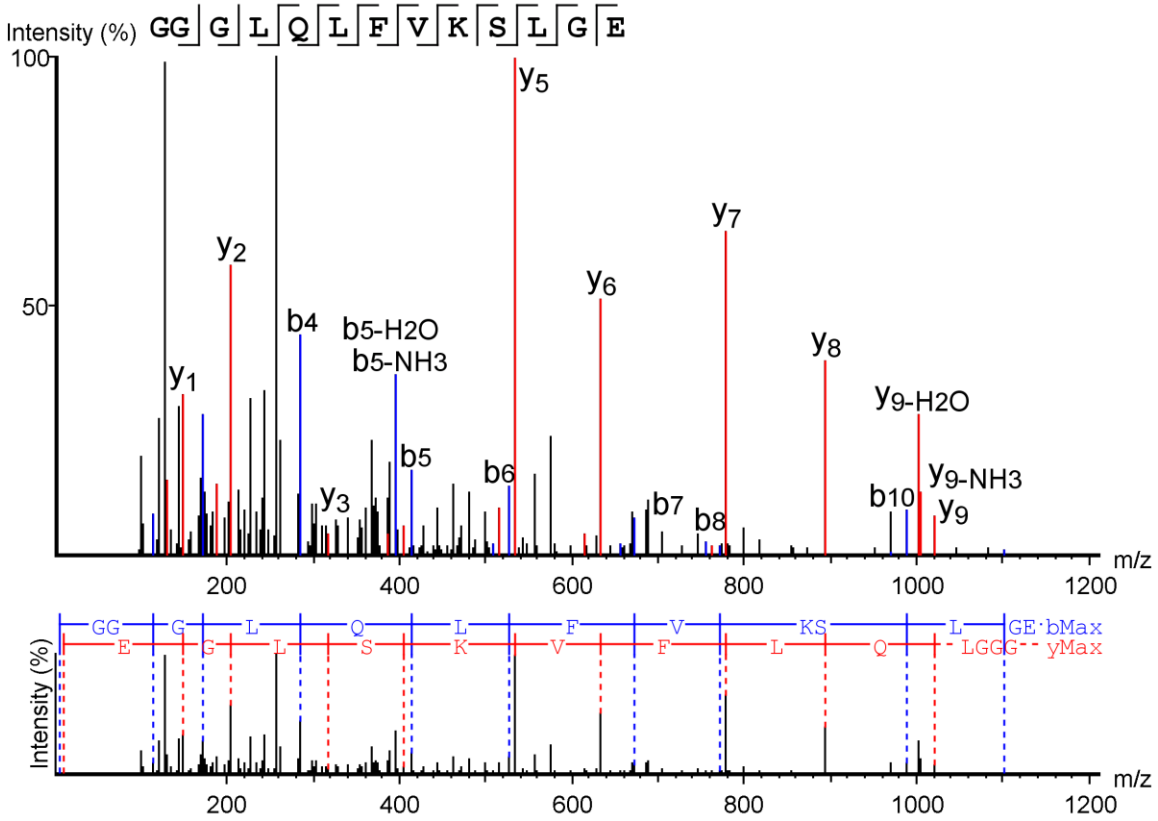
t6



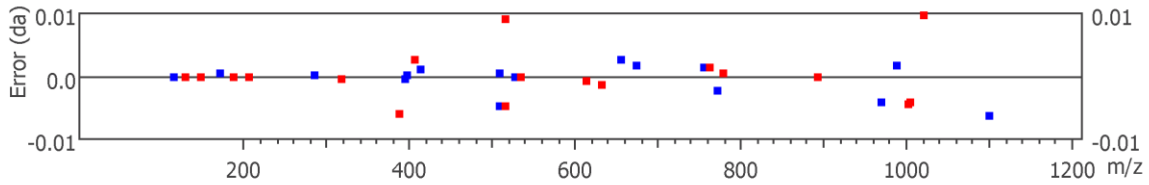
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					14
2	229.1183	211.1077	212.0918	115.0594	L	1607.7441	1589.7336	1590.7172	804.3721	13
3	316.1505	298.1398	299.1239	158.5754	S	1494.6594	1476.6497	1477.6332	747.8301	12
4	430.1938	412.1832	413.1668	215.5969	N	1407.6252	1389.6176	1390.6011	704.3140	11
5	545.2207	527.2101	528.1937	273.1104	D	1293.5831	1275.5747	1276.5582	647.2926	10
6	660.2477	642.2371	643.2207	330.6238	D	1178.5579	1160.5477	1161.5308	589.7791	9
7	830.3644	812.3538	813.3374	415.6822	R(+14.02)	1063.5312	1045.5208	1046.5027	532.2690	8
8	990.3951	972.3845	973.3681	495.6975	C(+57.02)	893.4146	875.4040	876.3876	447.2073	7
9	1119.4376	1101.4271	1102.4106	560.2188	E	733.3817	715.3733	716.3569	367.1919	6
10	1247.4962	1229.4857	1230.4692	624.2481	Q	604.3420	586.3307	587.3151	302.6707	5
11	1360.5803	1342.5698	1343.5533	680.7902	L	476.2827	458.2681	459.2557	238.6414	4
12	1516.6814	1498.6709	1499.6544	758.8407	R	363.1984	345.1881	346.1725	182.0993	3
13	1617.7291	1599.7186	1600.7021	809.3646	T	207.0978	189.0871	190.0705	104.0488	2
14					S	106.0499	88.0393	89.0229	53.5249	1



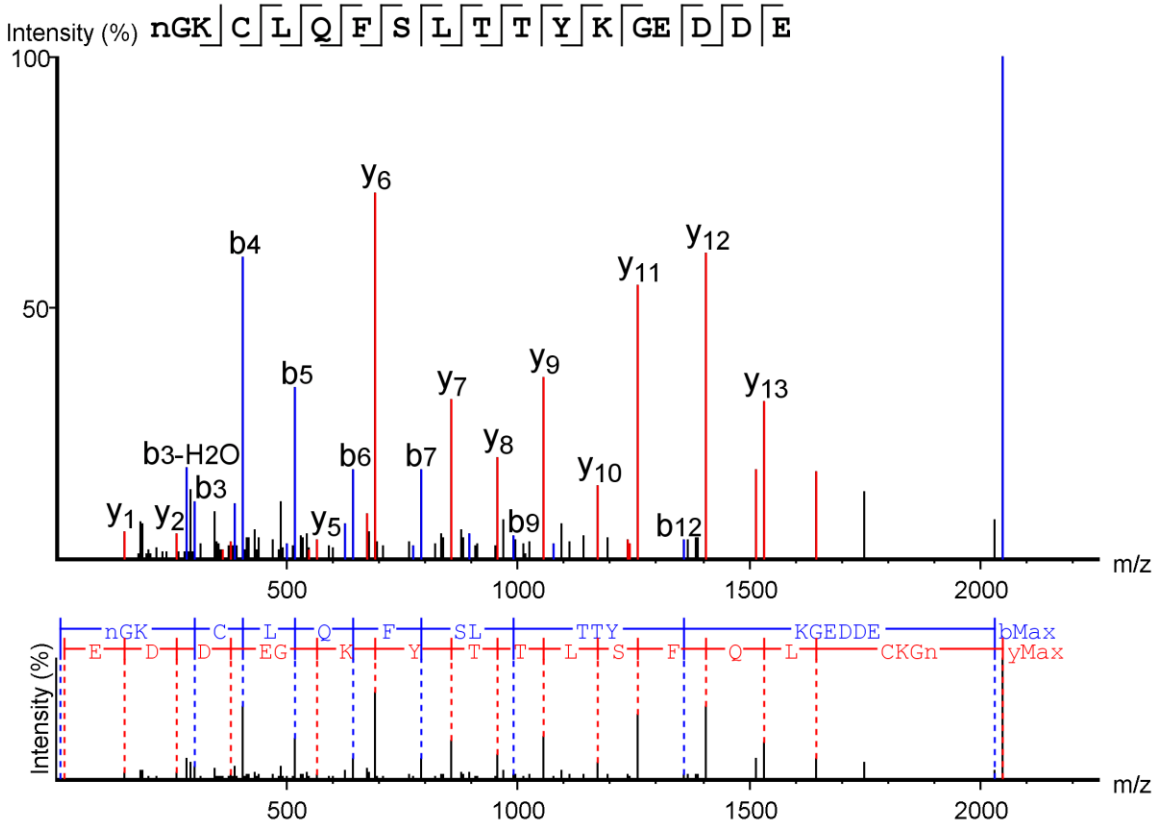
g1



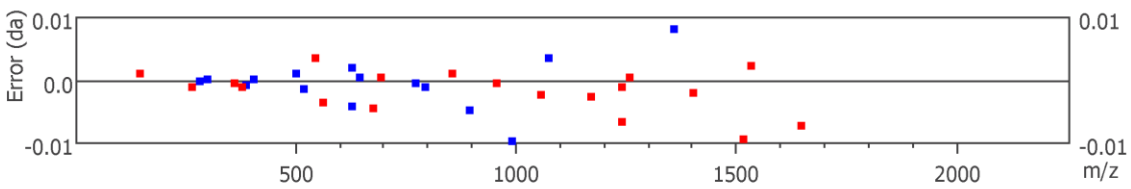
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					13
2	115.0505	97.0402	98.0238	58.0254	G	1247.6993	1229.6888	1230.6724	624.3497	12
3	172.0715	154.0617	155.0452	86.5361	G	1190.6779	1172.6674	1173.6509	595.8389	11
4	285.1558	267.1457	268.1293	143.0781	L	1133.6565	1115.6460	1116.6295	567.3282	10
5	413.2134	395.2047	396.1873	207.1074	Q	1020.5625	1002.5663	1003.5496	510.7862	9
6	526.2991	508.2875	509.2767	263.6494	L	892.5139	874.5032	875.4868	446.7569	8
7	673.3652	655.3539	656.3403	337.1837	F	779.4288	761.4175	762.4028	390.2149	7
8	772.4381	754.4235	755.4088	386.7179	V	632.3628	614.3514	615.3344	316.6807	6
9	900.5306	882.5201	883.5037	450.7653	K	533.2927	515.2728	516.2708	267.1465	5
10	987.5608	969.5521	970.5397	494.2813	S	405.1951	387.1934	388.1710	203.0990	4
11	1100.6530	1082.6362	1083.6198	550.8234	L	318.1663	300.1554	301.1389	159.5830	3
12	1157.6682	1139.6577	1140.6412	579.3341	G	205.0821	187.0714	188.0549	103.0409	2
13					E	148.0604	130.0500	131.0334	74.5302	1



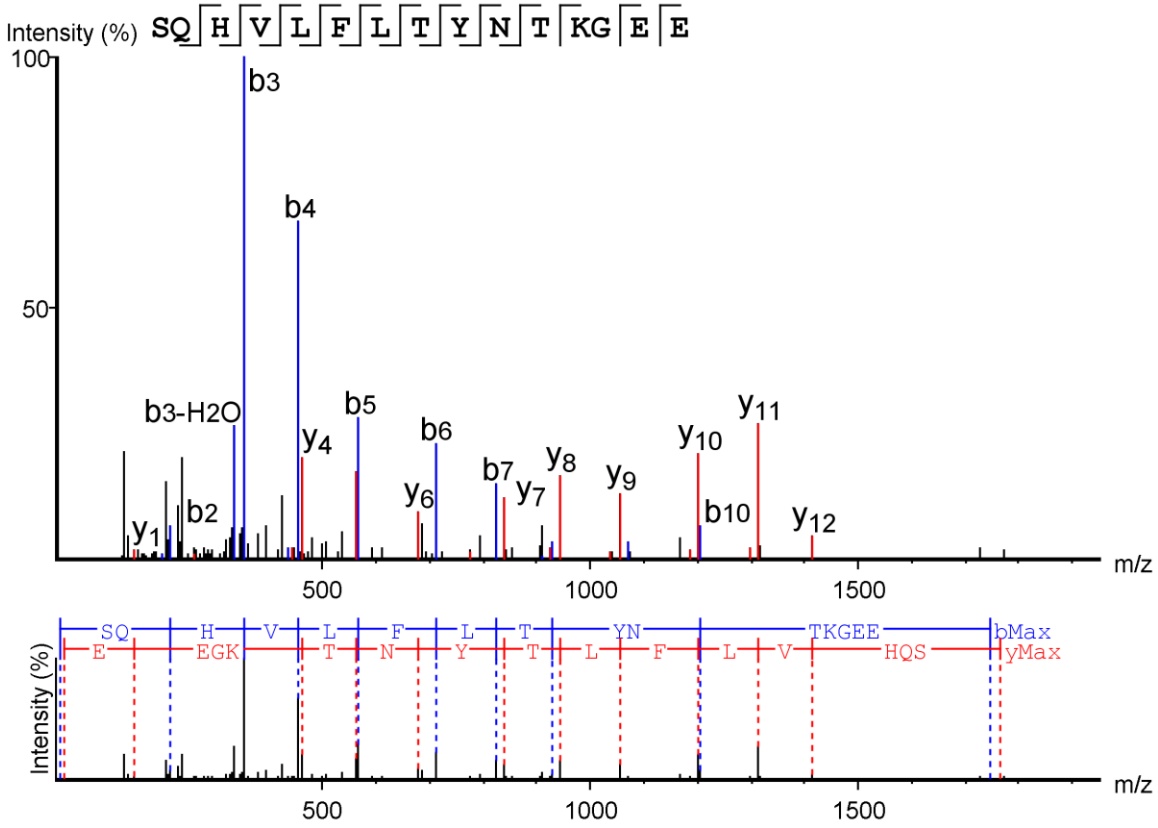
g2



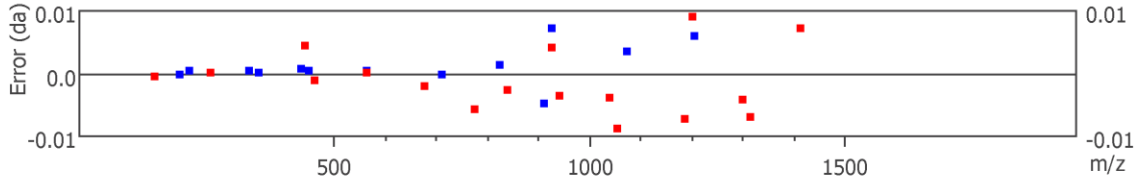
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	N(+.98)					18
2	173.0562	155.0457	156.0292	87.0281	G	1933.8848	1915.8743	1916.8578	967.4424	17
3	301.1508	283.1409	284.1242	151.0756	K	1876.8633	1858.8528	1859.8363	938.9316	16
4	404.1599	386.1505	387.1334	202.5802	C	1748.7683	1730.7578	1731.7413	874.8842	15
5	517.2458	499.2324	500.2175	259.1222	L	1645.7662	1627.7487	1628.7322	823.3796	14
6	645.3022	627.2900	628.2803	323.1515	Q	1532.6724	1514.6646	1515.6575	766.8375	13
7	792.3726	774.3612	775.3444	396.6857	F	1404.6184	1386.6060	1387.5895	702.8082	12
8	879.4034	861.3929	862.3765	440.2017	S	1257.5474	1239.5442	1240.5221	629.2740	11
9	992.4972	974.4769	975.4605	496.7437	L	1170.5187	1152.5055	1153.4890	585.7580	10
10	1093.5352	1075.5208	1076.5082	547.2676	T	1057.4343	1039.4215	1040.4050	529.2160	9
11	1194.5829	1176.5724	1177.5559	597.7914	T	956.3849	938.3737	939.3573	478.6921	8
12	1357.6377	1339.6357	1340.6193	679.3231	Y	855.3351	837.3260	838.3096	428.1683	7
13	1485.7412	1467.7307	1468.7142	743.3706	K	692.2726	674.2672	675.2463	346.6367	6
14	1542.7626	1524.7521	1525.7356	771.8813	G	564.1819	546.1641	547.1514	282.5892	5
15	1671.8052	1653.7947	1654.7782	836.4026	E	507.1569	489.1463	490.1299	254.0784	4
16	1786.8322	1768.8217	1769.8052	893.9209	D	378.1152	360.1042	361.0873	189.5571	3
17	1901.8591	1883.8486	1884.8322	951.4296	D	263.0884	245.0768	246.0603	132.0437	2
18					E	148.0590	130.0499	131.0334	74.5302	1



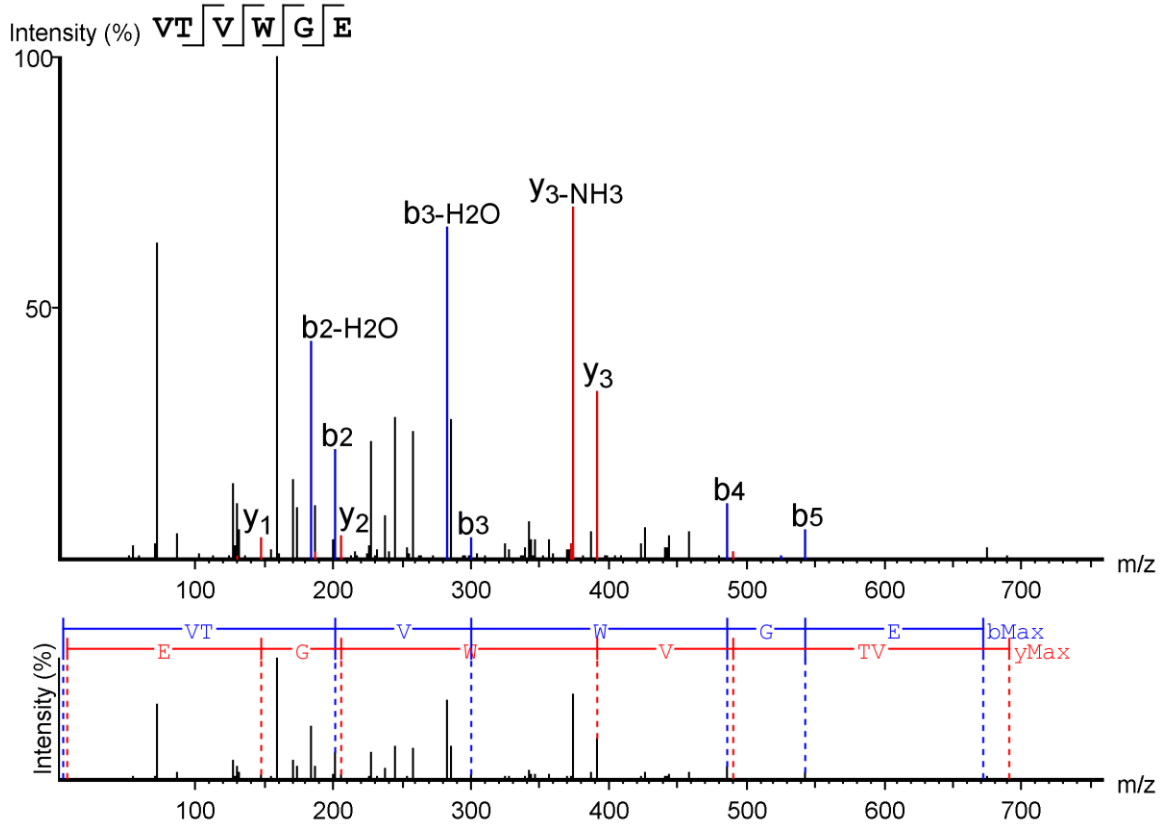
g3



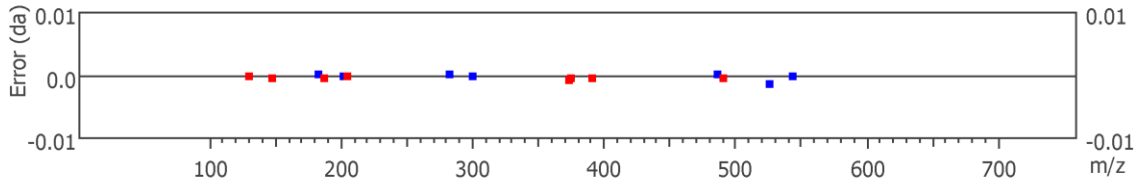
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	88.0399	70.0293	71.0129	44.5199	S					15
2	216.0976	198.0879	199.0713	108.5492	Q	1678.8435	1660.8330	1661.8165	839.9218	14
3	353.1567	335.1461	336.1304	177.0787	H	1550.7849	1532.7744	1533.7579	775.8981	13
4	452.2249	434.2152	435.1978	226.6129	V	1413.7183	1395.7155	1396.6990	707.3630	12
5	565.3091	547.2993	548.2828	283.1549	L	1314.6646	1296.6471	1297.6348	657.8288	11
6	712.3783	694.3677	695.3513	356.6891	F	1201.5640	1183.5630	1184.5537	601.2867	10
7	825.4604	807.4517	808.4353	413.2311	L	1054.5138	1036.4946	1037.4821	527.7526	9
8	926.5024	908.4994	909.4878	463.7550	T	941.4244	923.4105	924.3895	471.2105	8
9	1089.5732	1071.5588	1072.5463	545.2866	Y	840.3761	822.3628	823.3464	420.6867	7
10	1203.6099	1185.6057	1186.5892	602.3081	N	677.3121	659.2995	660.2831	339.1550	6
11	1304.6639	1286.6534	1287.6370	652.8320	T	563.2666	545.2565	546.2401	282.1335	5
12	1432.7588	1414.7483	1415.7318	716.8794	K	462.2204	444.2041	445.1924	231.6097	4
13	1489.7803	1471.7698	1472.7533	745.3901	G	334.1245	316.1139	317.0975	167.5622	3
14	1618.8229	1600.8124	1601.7959	809.9114	E	277.1030	259.0919	260.0760	139.0515	2
15					E	148.0607	130.0499	131.0334	74.5302	1



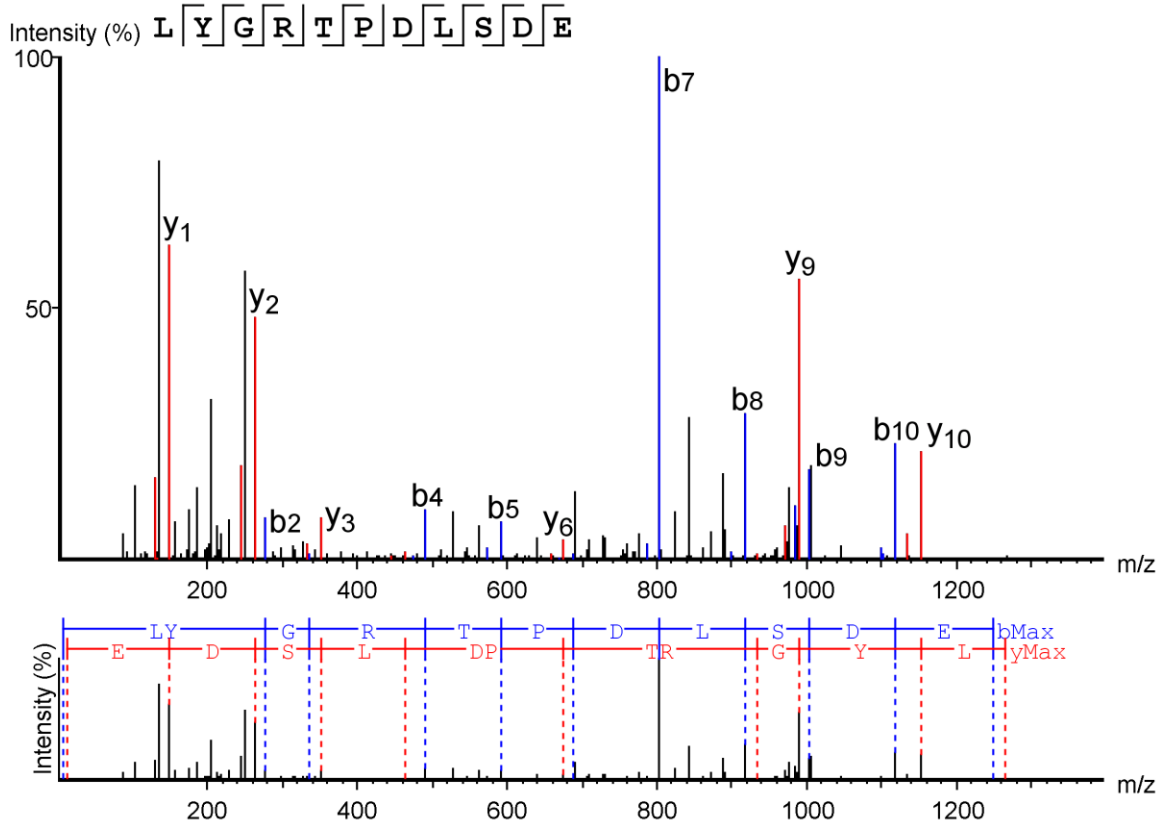
g4



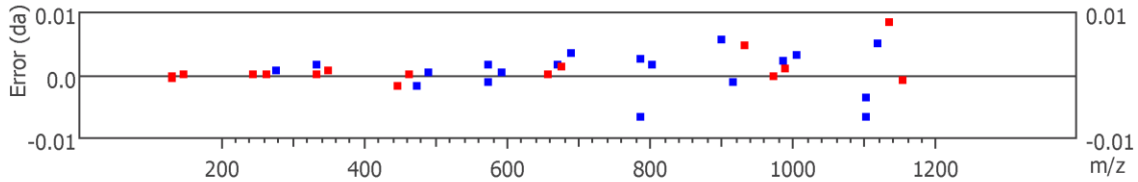
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	100.0762	82.0657	83.0492	50.5381	V					6
2	201.1236	183.1129	184.0969	101.0620	T	591.2773	573.2667	574.2503	296.1386	5
3	300.1922	282.1813	283.1653	150.5962	V	490.2302	472.2191	473.2026	245.6148	4
4	486.2711	468.2611	469.2447	243.6358	W	391.1616	373.1514	374.1348	196.0806	3
5	543.2930	525.2838	526.2661	272.1465	G	205.0820	187.0717	188.0549	103.0409	2
6					E	148.0607	130.0499	131.0334	74.5302	1



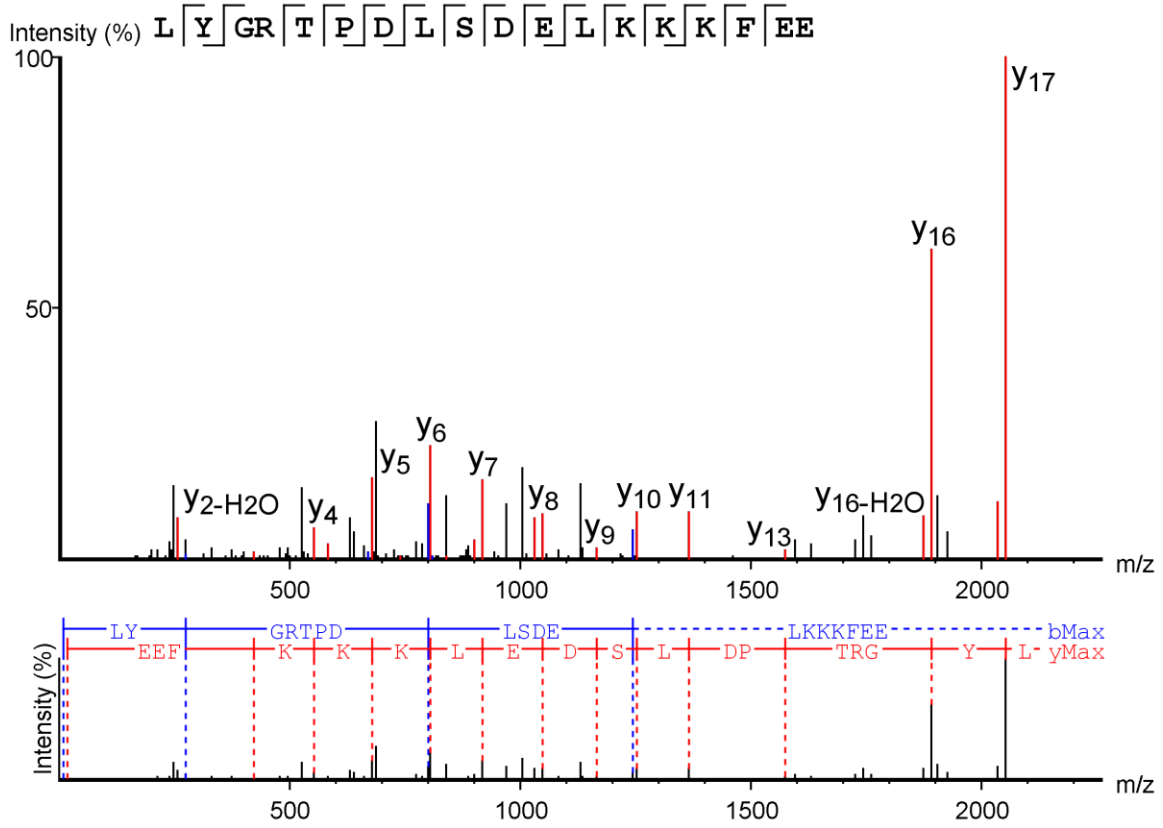
g5



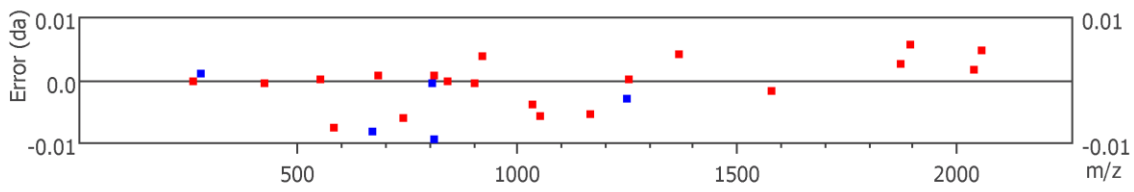
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					11
2	277.1541	259.1447	260.1282	139.0776	Y	1152.5176	1134.4976	1135.4897	576.7584	10
3	334.1747	316.1661	317.1497	167.5883	G	989.4519	971.4425	972.4264	495.2267	9
4	490.2770	472.2672	473.2525	245.6389	R	932.4268	914.4214	915.4050	466.7160	8
5	591.3246	573.3127	574.2997	296.1627	T	776.3308	758.3203	759.3038	388.6654	7
6	688.3743	670.3657	671.3512	344.6891	P	675.2816	657.2722	658.2562	338.1416	6
7	803.4031	785.3916	786.3849	402.2026	D	578.2303	560.2198	561.2034	289.6152	5
8	916.4903	898.4725	899.4622	458.7446	L	463.2031	445.1945	446.1765	232.1017	4
9	1003.5177	985.5079	986.4943	502.2606	S	350.1184	332.1085	333.0924	175.5597	3
10	1118.5428	1100.5413	1101.5277	559.7741	D	263.0869	245.0763	246.0603	132.0437	2
11					E	148.0601	130.0497	131.0339	74.5302	1



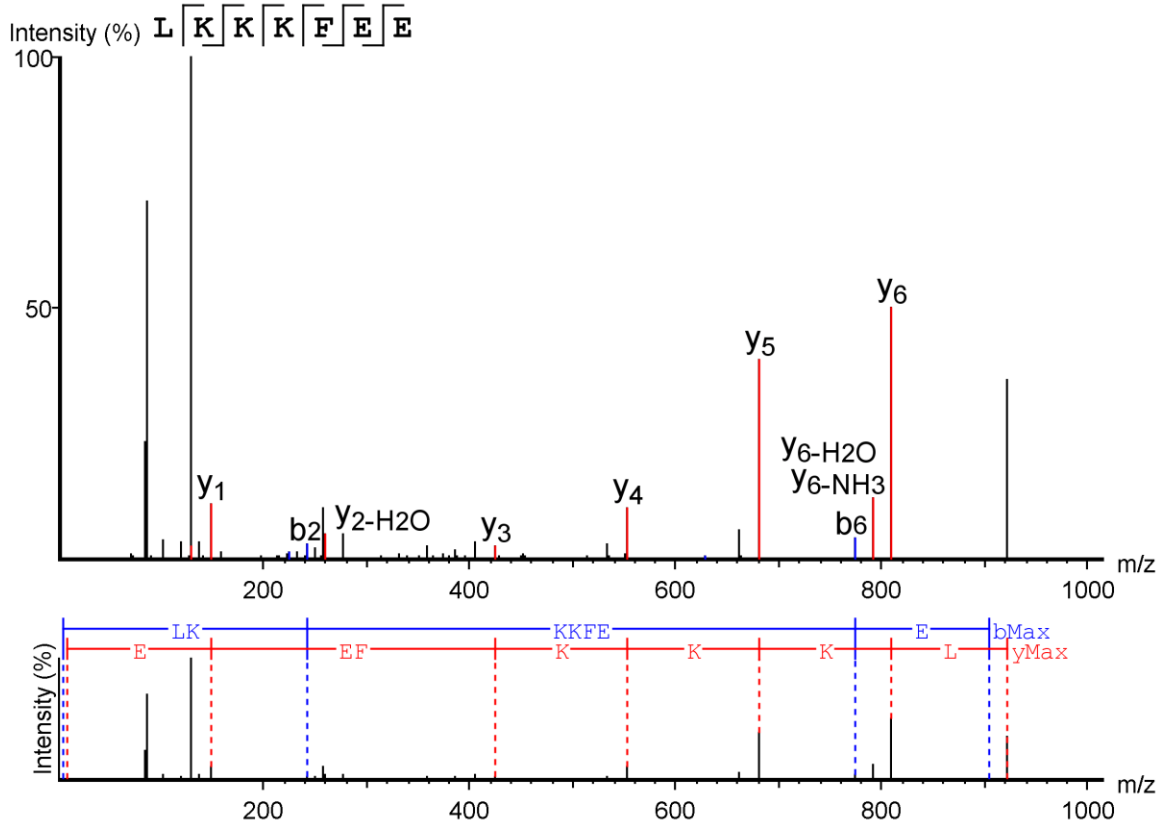
g6



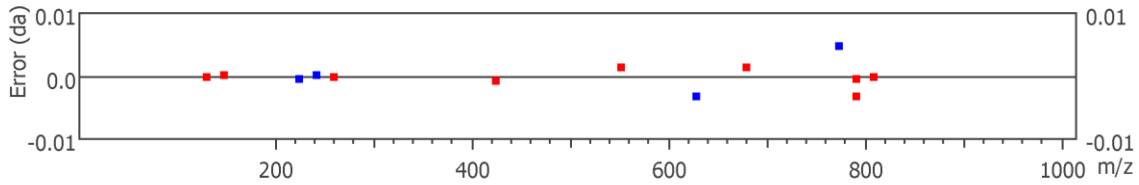
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					18
2	277.1537	259.1447	260.1282	139.0776	Y	2055.0342	2037.0267	2038.0122	1028.0197	17
3	334.1767	316.1661	317.1497	167.5883	G	1891.9700	1873.9625	1874.9489	946.4879	16
4	490.2778	472.2672	473.2508	245.6389	R	1834.9545	1816.9440	1817.9275	917.9772	15
5	591.3254	573.3149	574.2985	296.1627	T	1678.8533	1660.8428	1661.8263	839.9268	14
6	688.3782	670.3676	671.3593	344.6891	P	1577.8073	1559.7952	1560.7787	789.4028	13
7	803.4055	785.3946	786.3782	402.2026	D	1480.7529	1462.7424	1463.7260	740.8826	12
8	916.4892	898.4786	899.4622	458.7446	L	1365.7216	1347.7155	1348.6990	683.3630	11
9	1003.5212	985.5107	986.4943	502.2606	S	1252.6415	1234.6313	1235.6149	626.8209	10
10	1118.5482	1100.5377	1101.5212	559.7741	D	1165.6152	1147.5994	1148.5829	583.3125	9
11	1247.5938	1229.5803	1230.5638	624.2954	E	1050.5885	1032.5762	1033.5559	525.7914	8
12	1360.6748	1342.6643	1343.6478	680.8374	L	921.5363	903.5304	904.5134	461.2702	7
13	1488.7698	1470.7593	1471.7428	744.8849	K	808.4551	790.4457	791.4293	404.7281	6
14	1616.8647	1598.8542	1599.8378	808.9417	K	680.3601	662.3508	663.3344	340.6807	5
15	1744.9597	1726.9492	1727.9327	872.9799	K	552.2659	534.2558	535.2394	276.6332	4
16	1892.0281	1874.0176	1875.0011	946.5140	F	424.1719	406.1608	407.1444	212.5857	3
17	2021.0707	2003.0602	2004.0437	1011.0353	E	277.1030	259.0924	260.0760	139.0515	2
18					E	148.0604	130.0499	131.0334	74.5302	1



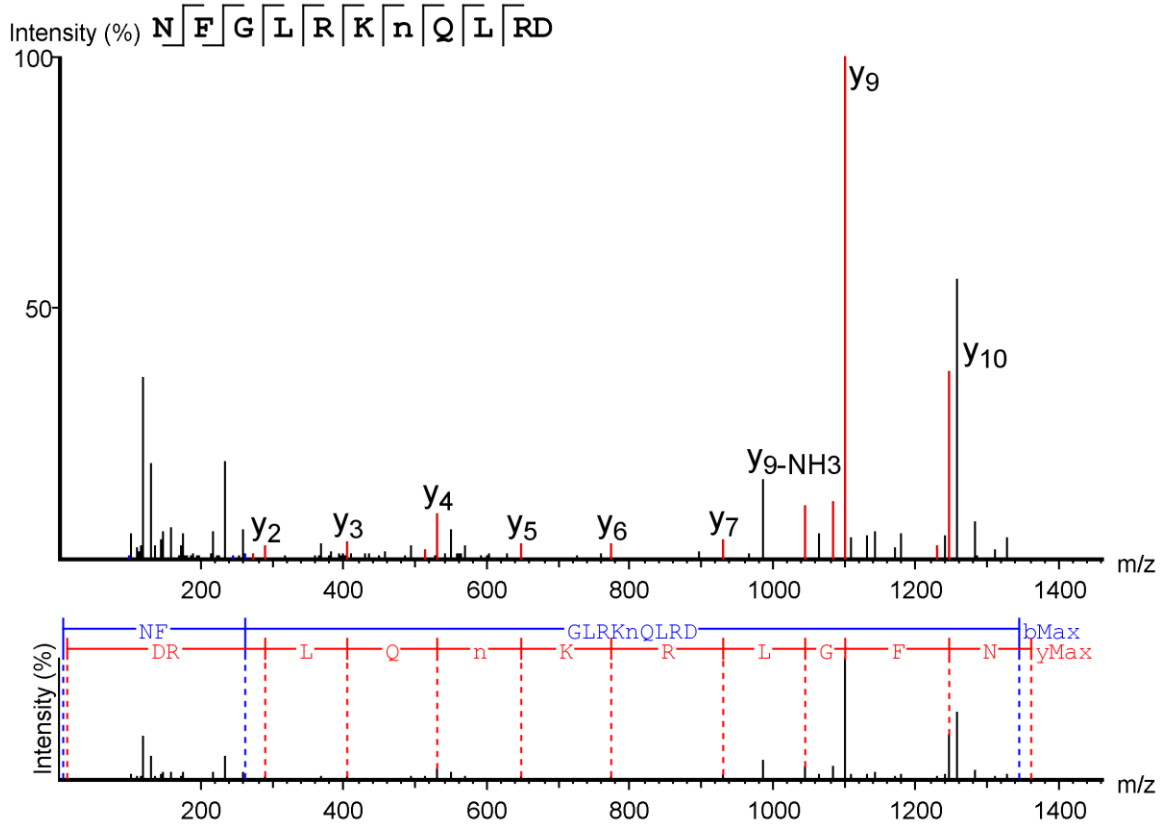
g7



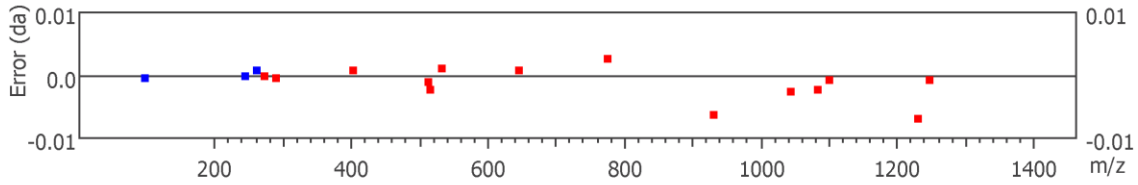
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	114.0919	96.0813	97.0649	57.5459	L					7
2	242.1865	224.1763	225.1603	121.5934	K	808.4564	790.4489	791.4297	404.7281	6
3	370.2818	352.2712	353.2548	185.6409	K	680.3596	662.3508	663.3344	340.6807	5
4	498.3767	480.3662	481.3498	249.6884	K	552.2645	534.2558	535.2394	276.6332	4
5	645.4451	627.4377	628.4182	323.2226	F	424.1720	406.1608	407.1444	212.5857	3
6	774.4827	756.4772	757.4608	387.7439	E	277.1030	259.0927	260.0760	139.0515	2
7					E	148.0601	130.0499	131.0334	74.5302	1



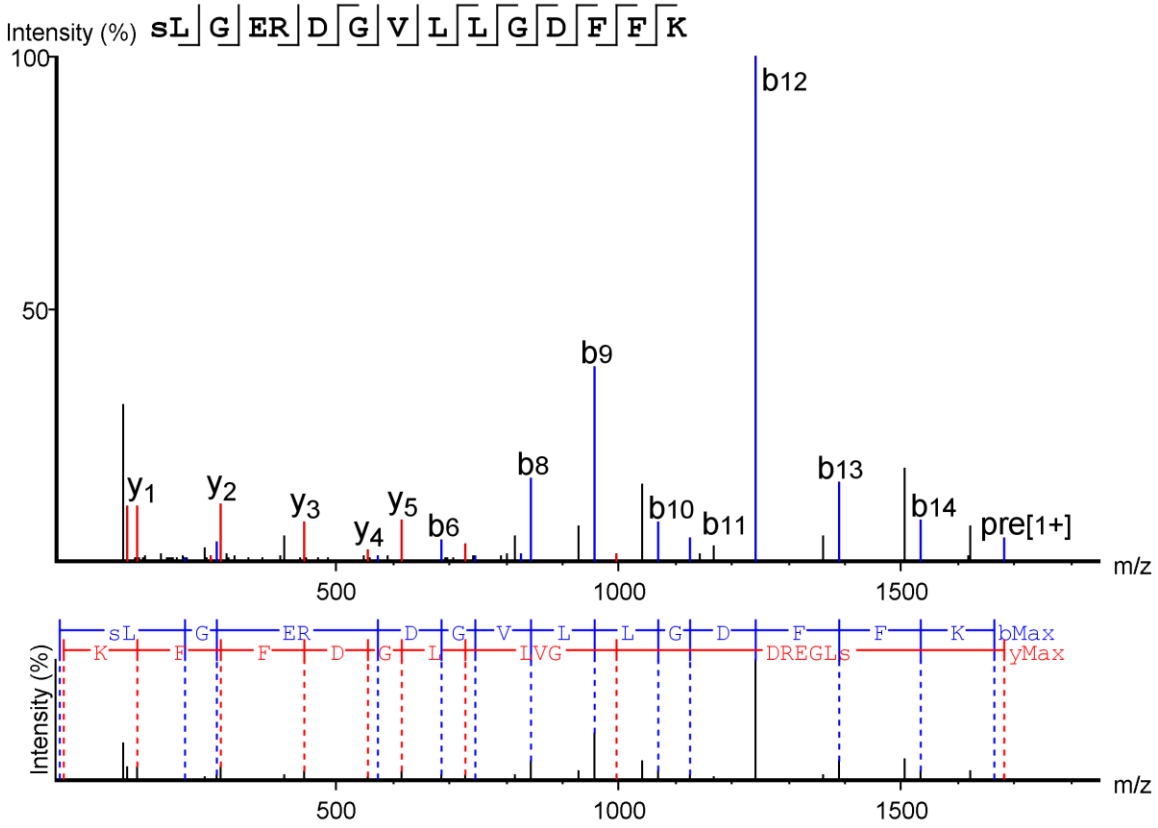
g8



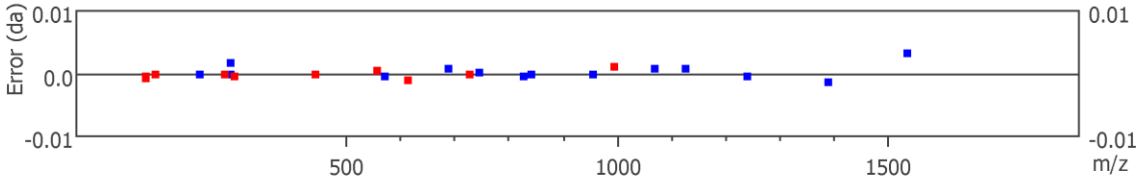
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	115.0508	97.0402	98.0241	58.0254	N					11
2	262.1179	244.1086	245.0924	131.5596	F	1247.6862	1229.6749	1230.6653	624.3427	10
3	319.1406	301.1301	302.1136	160.0703	G	1100.6179	1082.6066	1083.5925	550.8085	9
4	432.2247	414.2141	415.1977	216.6123	L	1043.5981	1025.5851	1026.5686	522.2978	8
5	588.3258	570.3152	571.2988	294.6629	R	930.5177	912.5010	913.4846	465.7558	7
6	716.4207	698.4102	699.3937	358.7104	K	774.4075	756.3998	757.3834	387.7052	6
7	831.4477	813.4371	814.4207	416.2238	N(+.98)	646.3145	628.3049	629.2885	323.6577	5
8	959.5062	941.4957	942.4792	480.2531	Q	531.2872	513.2791	514.2638	266.1443	4
9	1072.5903	1054.5798	1055.5634	536.7952	L	403.2287	385.2194	386.2029	202.1150	3
10	1228.6914	1210.6809	1211.6644	614.8457	R	290.1463	272.1353	273.1191	145.5729	2
11					D	134.0448	116.0342	117.0178	67.5224	1



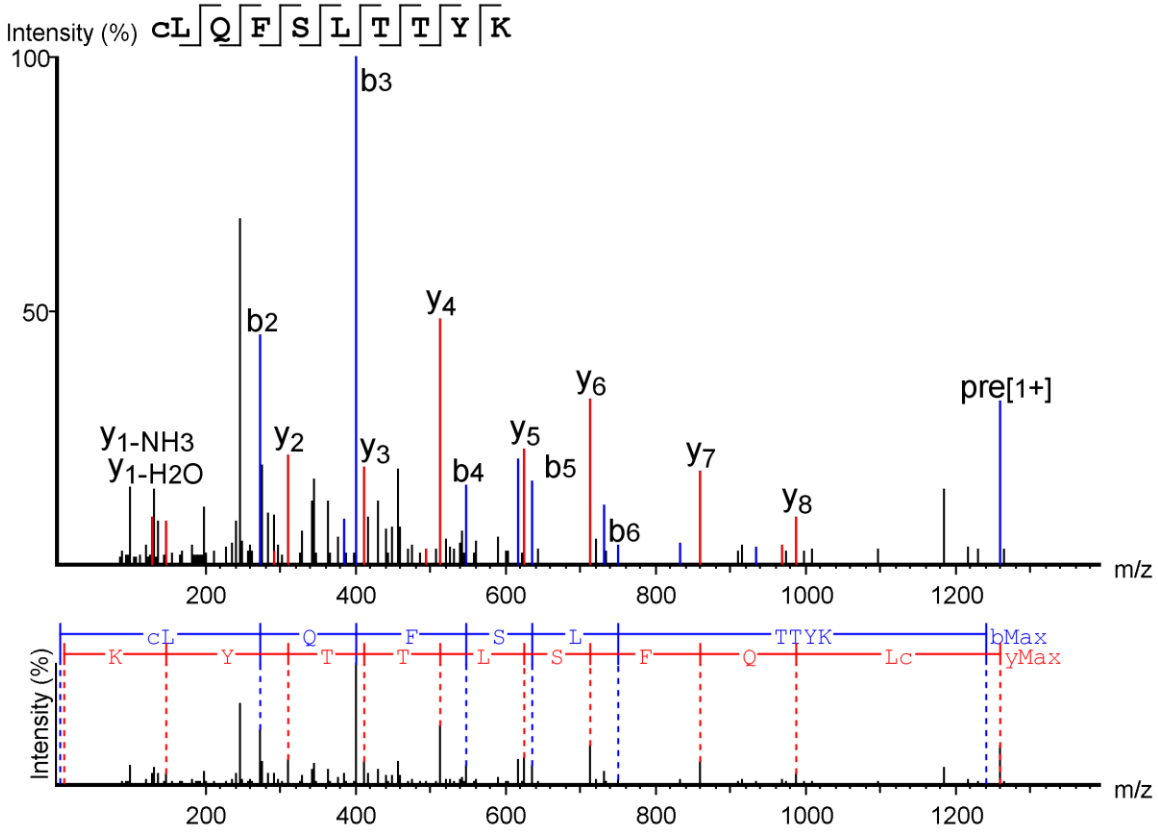
I1



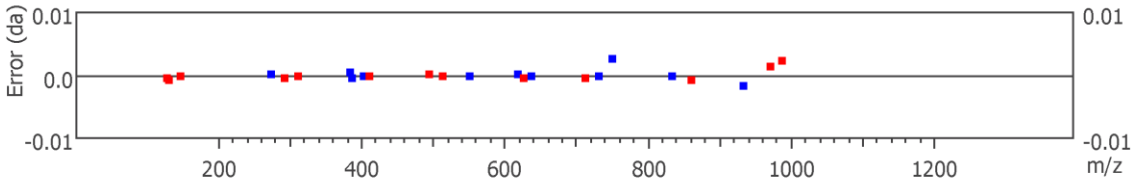
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	S(+27.99)					15
2	229.1186	211.1083	212.0918	115.0594	L	1565.8322	1547.8217	1548.8052	783.4161	14
3	286.1401	268.1297	269.1133	143.5701	G	1452.7480	1434.7375	1435.7211	726.8740	13
4	415.1829	397.1723	398.1559	208.0914	E	1395.7266	1377.7161	1378.6996	698.3633	12
5	571.2844	553.2734	554.2570	286.1401	R	1266.6841	1248.6736	1249.6571	633.8420	11
6	686.3099	668.3004	669.2839	343.6555	D	1110.5829	1092.5724	1093.5559	555.7914	10
7	743.3318	725.3218	726.3054	372.1662	G	995.5544	977.5454	978.5290	498.2780	9
8	842.4008	824.3902	825.3744	421.7004	V	938.5345	920.5240	921.5076	469.7673	8
9	955.4851	937.4743	938.4579	478.2424	L	839.4661	821.4556	822.4391	420.2331	7
10	1068.5676	1050.5583	1051.5419	534.7844	L	726.3824	708.3715	709.3551	363.6910	6
11	1125.5894	1107.5798	1108.5634	563.2952	G	613.2990	595.2874	596.2710	307.1490	5
12	1240.6179	1222.6068	1223.5903	620.8087	D	556.2758	538.2660	539.2496	278.6383	4
13	1387.6870	1369.6752	1370.6587	694.3428	F	441.2499	423.2390	424.2226	221.1248	3
14	1534.7506	1516.7437	1517.7272	767.8771	F	294.1816	276.1708	277.1542	147.5906	2
15					K	147.1130	129.1027	130.0866	74.0564	1



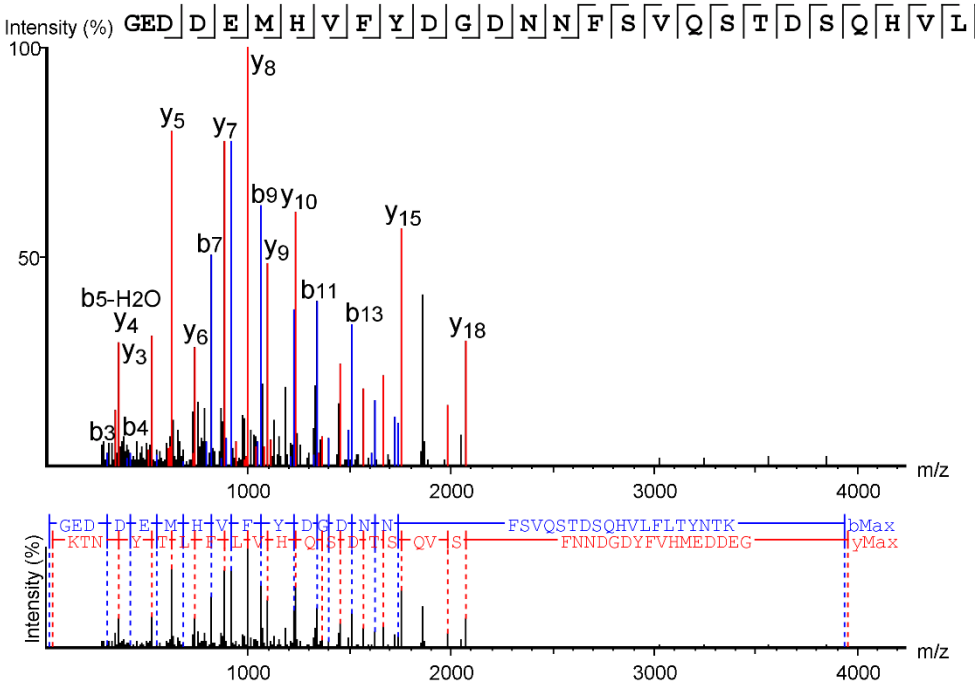
I2



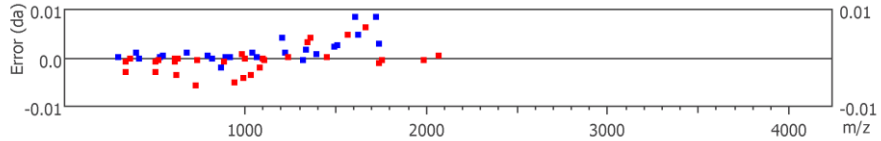
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	161.0385	143.0279	144.0115	81.0192	C(+57.02)					10
2	274.1222	256.1120	257.0956	137.5613	L	1100.5986	1082.5881	1083.5717	550.7993	9
3	402.1809	384.1697	385.1546	201.5905	Q	987.5120	969.5040	970.4857	494.2573	8
4	549.2495	531.2390	532.2225	275.1248	F	859.4568	841.4454	842.4290	430.2280	7
5	636.2816	618.2705	619.2546	318.6408	S	712.3881	694.3770	695.3606	356.6938	6
6	749.3625	731.3549	732.3386	375.1828	L	625.3561	607.3450	608.3286	313.1778	5
7	850.4133	832.4026	833.3863	425.7066	T	512.2716	494.2605	495.2445	256.6357	4
8	951.4609	933.4520	934.4340	476.2305	T	411.2236	393.2132	394.1968	206.1119	3
9	1114.5243	1096.5138	1097.4973	557.7621	Y	310.1763	292.1660	293.1491	155.5880	2
10					K	147.1129	129.1026	130.0865	74.0564	1



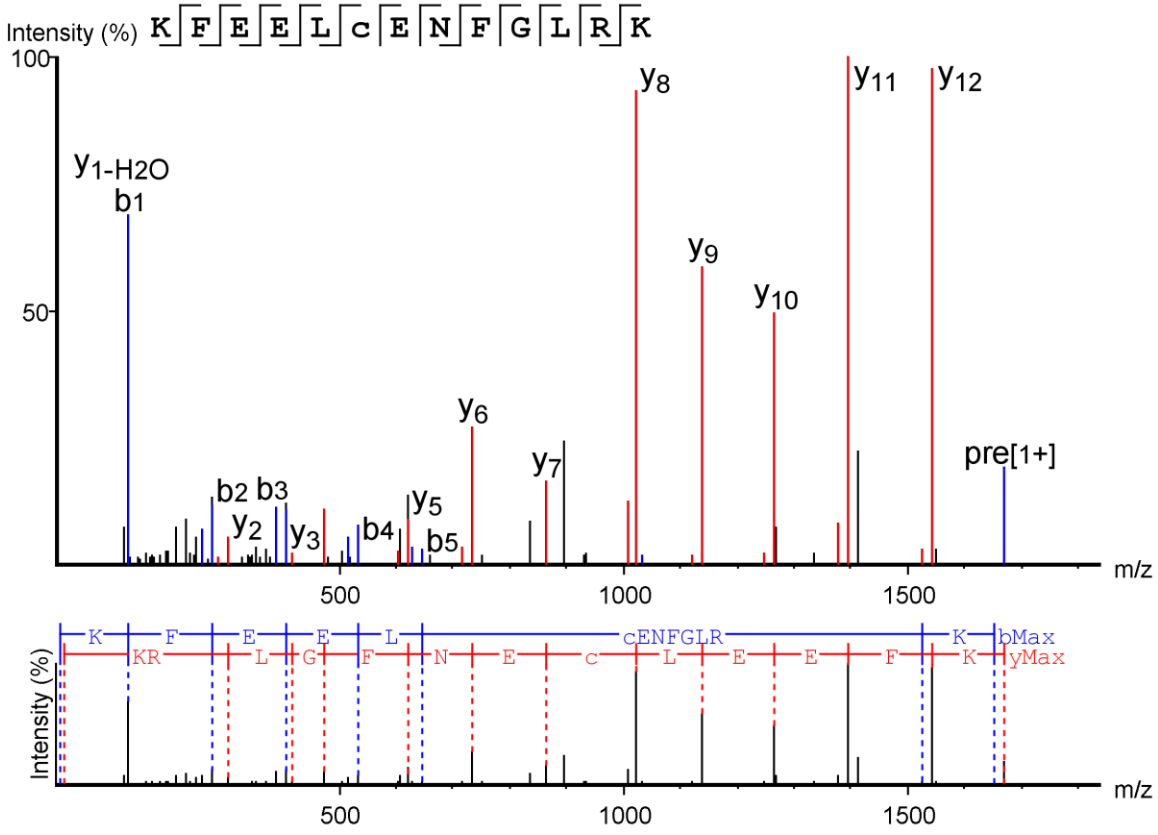
I3



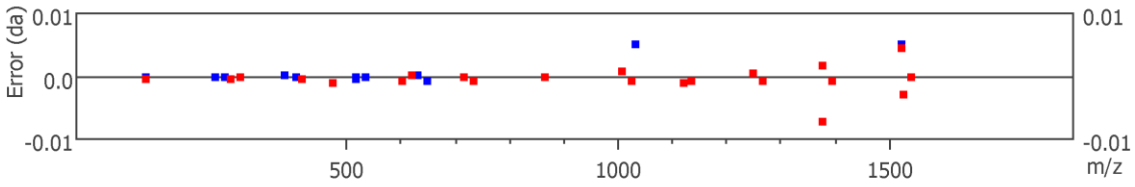
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	58.0293	40.0187	41.0023	29.5146	G					34
2	187.0719	169.0613	170.0449	94.0359	E	3895.7026	3877.6921	3878.6755	1948.3513	33
3	302.0985	284.0883	285.0718	151.5494	D	3766.6602	3748.6497	3749.6331	1883.8301	32
4	417.1259	399.1139	400.0988	209.0629	D	3651.6333	3633.6228	3634.6062	1826.3167	31
5	546.1676	528.1573	529.1414	273.5842	E	3536.6062	3518.5957	3519.5791	1768.8031	30
6	677.2073	659.1982	660.1818	339.1044	M	3407.5637	3389.5532	3390.5366	1704.2819	29
7	814.2678	796.2563	797.2408	407.6339	H	3276.5232	3258.5127	3259.4961	1638.7616	28
8	913.3358	895.3250	896.3092	457.1681	V	3139.4644	3121.4539	3122.4373	1570.2322	27
9	1060.4039	1042.3926	1043.3776	530.7023	F	3040.3958	3022.3853	3023.3687	1520.6979	26
10	1223.4664	1205.4529	1206.4409	612.2339	Y	2893.3274	2875.3169	2876.3003	1447.1637	25
11	1338.4929	1320.4849	1321.4679	669.7474	D	2730.2642	2712.2537	2713.2371	1365.6321	24
12	1395.5153	1377.5057	1378.4893	698.2581	G	2615.2373	2597.2268	2598.2102	1308.1187	23
13	1510.5403	1492.5300	1493.5162	755.7716	D	2558.2158	2540.2053	2541.1887	1279.6079	22
14	1624.5812	1606.5670	1607.5592	812.7931	N	2443.1887	2425.1782	2426.1616	1222.0944	21
15	1738.6259	1720.6185	1721.5933	869.8166	N	2329.1458	2311.1353	2312.1187	1165.0729	20
16	1885.6975	1867.6870	1868.6705	943.3488	F	2215.1030	2197.0925	2198.0759	1108.0521	19
17	1972.7295	1954.7190	1955.7025	986.8647	S	2068.0337	2050.0239	2051.0073	1034.5208	18
18	2071.7979	2053.7874	2054.7708	1036.3989	V	1981.0029	1962.9919	1963.9755	991.0054	17
19	2199.8564	2181.8459	2182.8293	1100.4282	Q	1881.9341	1863.9236	1864.9071	941.4720	16
20	2286.8884	2268.8779	2269.8613	1143.9442	S	1753.8761	1735.8650	1736.8496	877.4377	15
21	2387.9363	2369.9258	2370.9092	1194.4681	T	1666.8368	1648.8330	1649.8165	833.9218	14
22	2502.9631	2484.9526	2485.9360	1251.9816	D	1565.7908	1547.7853	1548.7688	783.3979	13
23	2589.9951	2571.9846	2572.9680	1295.4976	S	1450.7683	1432.7583	1433.7418	725.8901	12
24	2718.0537	2700.0432	2701.0266	1359.5269	Q	1363.7324	1345.7263	1346.7064	682.3684	11
25	2855.1128	2837.1023	2838.0857	1428.0564	H	1235.6776	1217.6677	1218.6512	618.3427	10
26	2954.1812	2936.1707	2937.1541	1477.5906	V	1098.6194	1080.6089	1081.5945	549.8097	9
27	3067.2651	3049.2546	3050.2380	1534.1326	L	999.5507	981.5393	982.5239	500.2755	8
28	3214.3335	3196.3230	3197.3064	1607.6667	F	886.4675	868.4563	869.4399	443.7334	7
29	3327.4177	3309.4072	3310.3906	1664.2089	L	739.3990	721.3879	722.3715	370.1992	6
30	3428.4653	3410.4548	3411.4382	1714.7327	T	626.3147	608.3046	609.2874	313.6572	5
31	3591.5286	3573.5181	3574.5015	1796.2643	Y	525.2672	507.2591	508.2405	263.1334	4
32	3705.5715	3687.5610	3688.5444	1853.2858	N	362.2036	344.1959	345.1771	181.6017	3
33	3806.6191	3788.6086	3789.5920	1903.8096	T	248.1605	230.1499	231.1335	124.5802	2
34					K	147.1128	129.1022	130.0858	74.0564	1



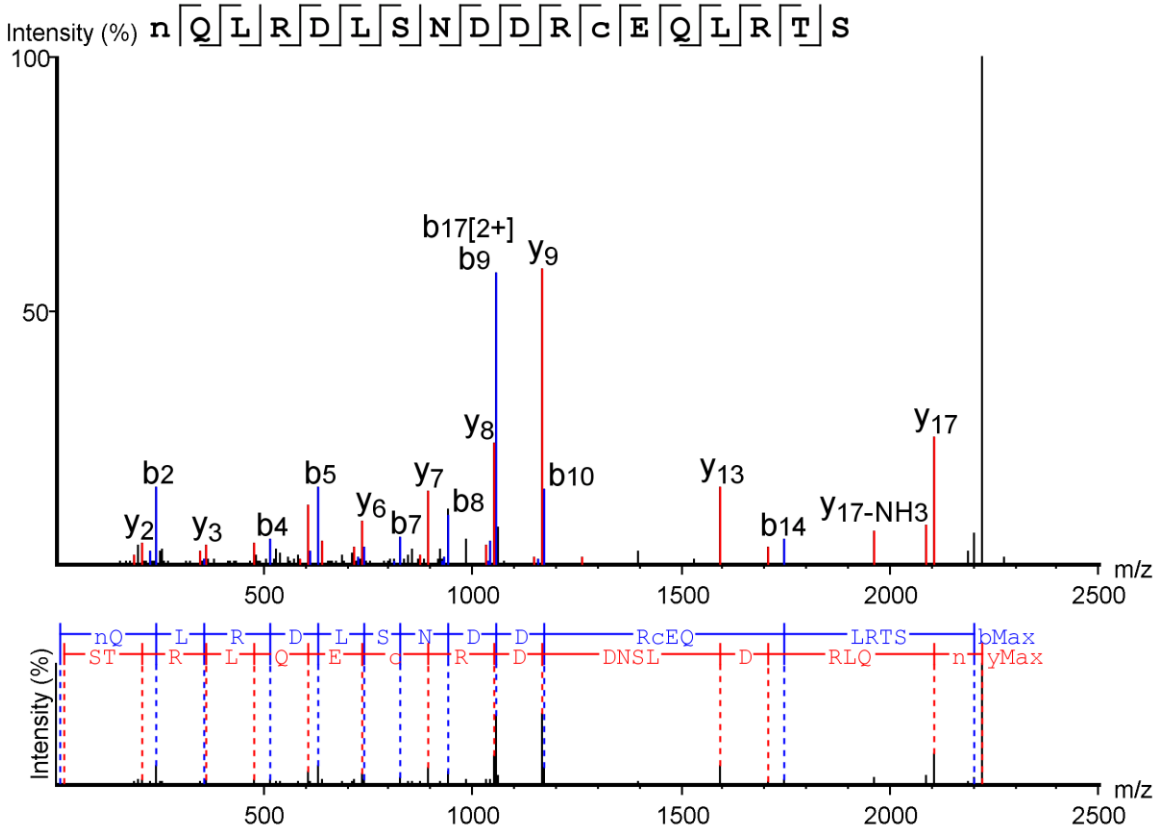
I4



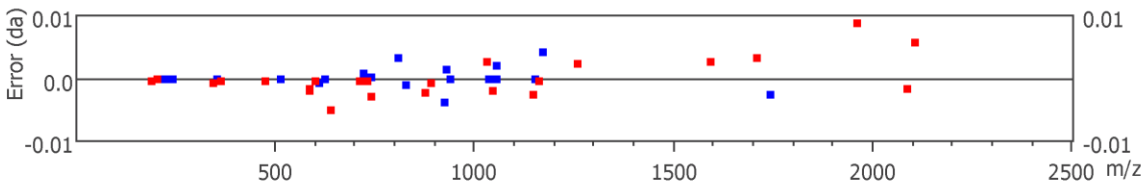
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	129.1026	111.0922	112.0758	65.0514	K					13
2	276.1713	258.1605	259.1445	138.5856	F	1541.7417	1523.7263	1524.7174	771.3708	12
3	405.2138	387.2027	388.1868	203.1069	E	1394.6741	1376.6606	1377.6536	697.8366	11
4	534.2564	516.2462	517.2294	267.6282	E	1265.6313	1247.6194	1248.6036	633.3153	10
5	647.3411	629.3292	630.3134	324.1702	L	1136.5889	1118.5775	1119.5623	568.7940	9
6	807.3711	789.3605	790.3441	404.1855	C(+57.02)	1023.5048	1005.4934	1006.4758	512.2520	8
7	936.4136	918.4031	919.3867	468.7068	E	863.4735	845.4628	846.4464	432.2367	7
8	1050.4565	1032.4460	1033.4243	525.7283	N	734.4314	716.4202	717.4034	367.7154	6
9	1197.5250	1179.5145	1180.4980	599.2625	F	620.3873	602.3773	603.3617	310.6939	5
10	1254.5464	1236.5359	1237.5194	627.7732	G	473.3205	455.3088	456.2924	237.1597	4
11	1367.6305	1349.6200	1350.6035	684.3152	L	416.2984	398.2874	399.2710	208.6490	3
12	1523.7263	1505.7211	1506.7046	762.3658	R	303.2141	285.2033	286.1874	152.1069	2
13					K	147.1128	129.1026	130.0858	74.0564	1



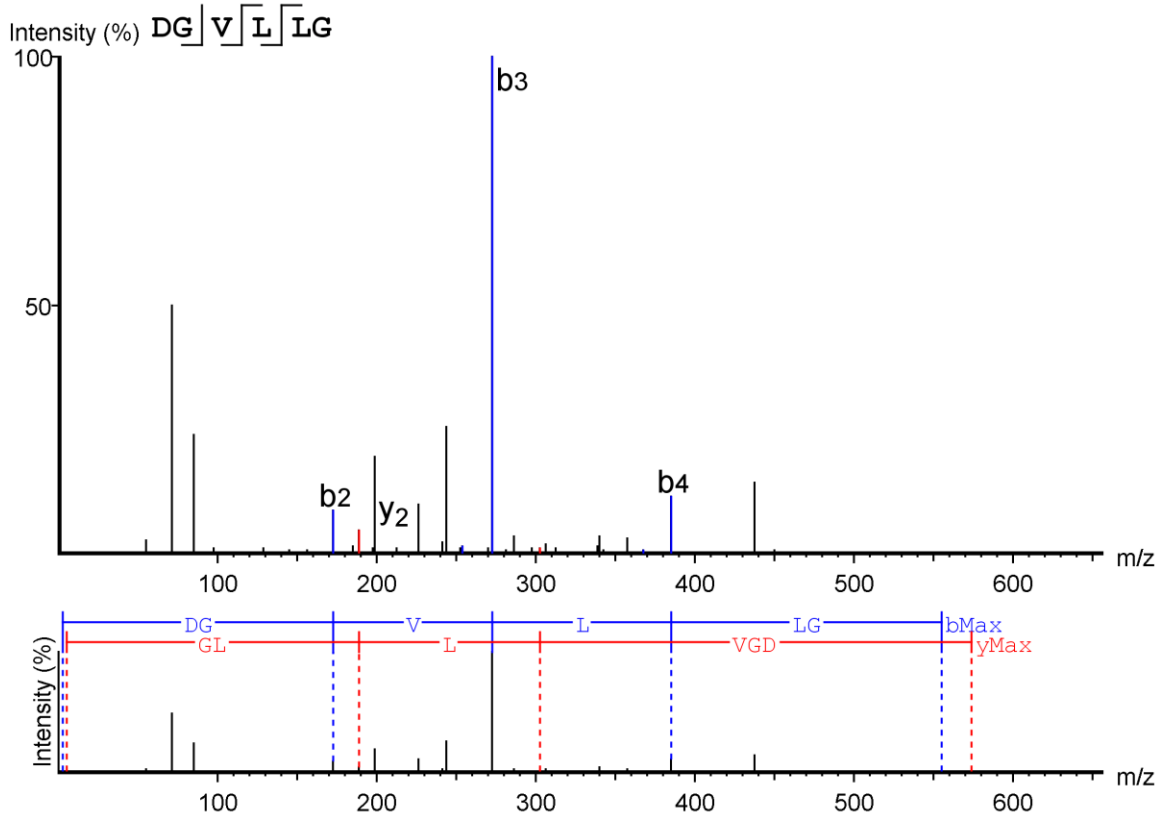
I5



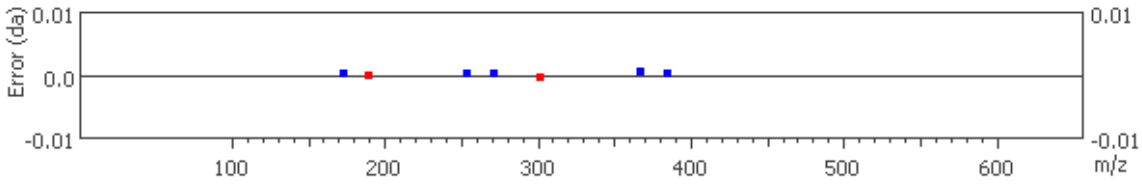
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	N(+.98)					18
2	244.0931	226.0826	227.0666	122.5467	Q	2105.9934	2087.9888	2088.9739	1053.4996	17
3	357.1772	339.1668	340.1504	179.0887	L	1977.9407	1959.9302	1960.9047	989.4703	16
4	513.2783	495.2679	496.2515	257.1393	R	1864.8566	1846.8461	1847.8296	932.9283	15
5	628.3052	610.2949	611.2792	314.6527	D	1708.7521	1690.7450	1691.7285	854.8777	14
6	741.3892	723.3790	724.3614	371.1948	L	1593.7256	1575.7180	1576.7015	797.3643	13
7	828.4225	810.4075	811.3945	414.7108	S	1480.6444	1462.6339	1463.6174	740.8251	12
8	942.4648	924.4539	925.4413	471.7322	N	1393.6124	1375.6019	1376.5854	697.3062	11
9	1057.4912	1039.4812	1040.4644	529.2457	D	1279.5695	1261.5590	1262.5399	640.2897	10
10	1172.5138	1154.5078	1155.4912	586.7592	D	1164.5431	1146.5321	1147.5182	582.7713	9
11	1328.6194	1310.6089	1311.5924	664.8097	R	1049.5176	1031.5051	1032.4858	525.2578	8
12	1488.6501	1470.6396	1471.6232	744.8251	C(+57.02)	893.4153	875.4040	876.3898	447.2073	7
13	1617.6926	1599.6821	1600.6656	809.3463	E	733.3844	715.3737	716.3569	367.1919	6
14	1745.7538	1727.7407	1728.7242	873.3756	Q	604.3419	586.3324	587.3163	302.6707	5
15	1858.8353	1840.8248	1841.8083	929.9161	L	476.2833	458.2722	459.2557	238.6414	4
16	2014.9364	1996.9259	1997.9094	1007.9682	R	363.1990	345.1881	346.1724	182.0993	3
17	2115.9841	2097.9736	2098.9570	1058.4896	T	207.0977	189.0874	190.0705	104.0488	2
18					S	106.0499	88.0393	89.0229	53.5249	1



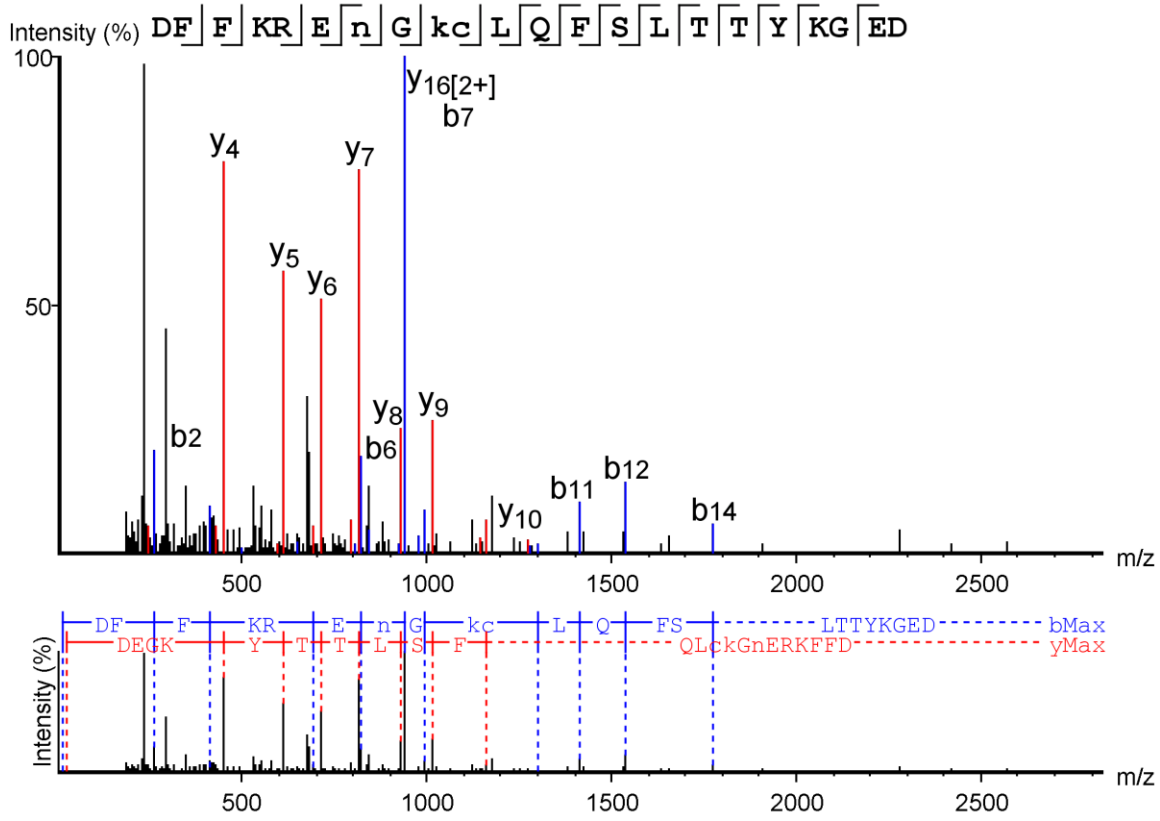
a1



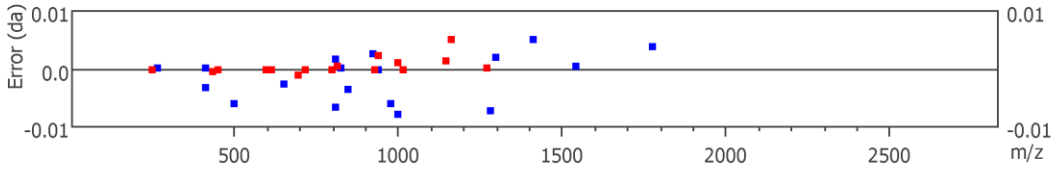
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					6
2	173.0559	155.0457	156.0292	87.0281	G	458.2973	440.2867	441.2703	229.6486	5
3	272.1241	254.1135	255.0977	136.5623	V	401.2758	383.2653	384.2488	201.1379	4
4	385.2083	367.1974	368.1817	193.1043	L	302.2078	284.1968	285.1804	151.6037	3
5	498.2928	480.2822	481.2658	249.6464	L	189.1237	171.1128	172.0964	95.0617	2
6					G	76.0393	58.0287	59.0123	38.5196	1



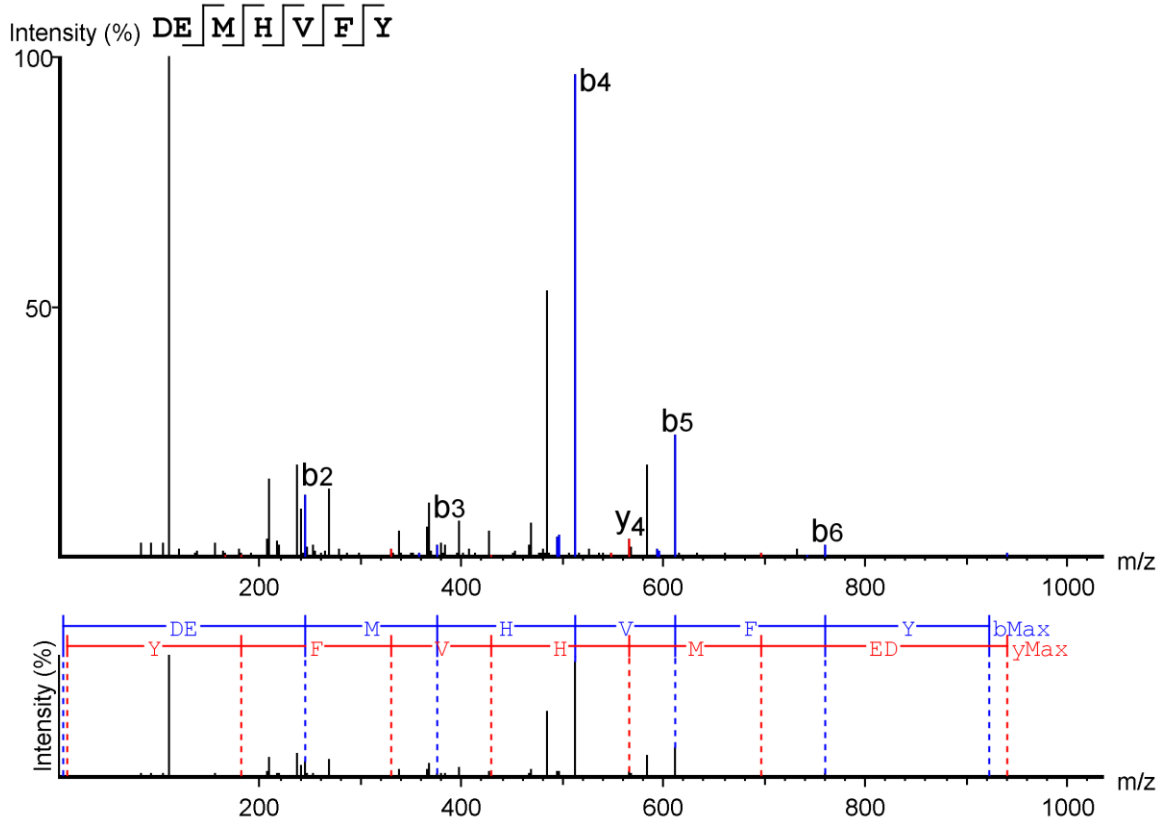
a2



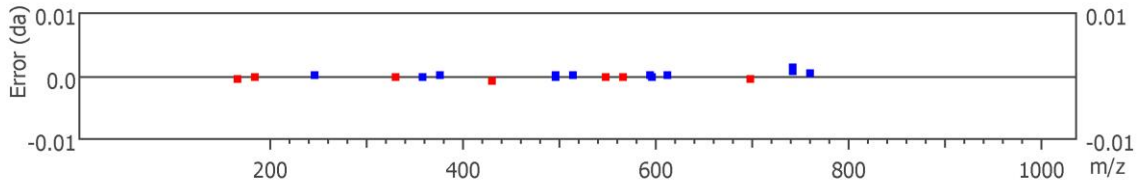
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					22
2	263.1027	245.0926	246.0762	132.0516	F	2583.2546	2565.2441	2566.2275	1292.1273	21
3	410.1710	392.1610	393.1446	205.5858	F	2436.1863	2418.1758	2419.1592	1218.5931	20
4	538.2665	520.2560	521.2396	269.6333	K	2289.1179	2271.1074	2272.0908	1145.0590	19
5	694.3688	676.3571	677.3406	347.6838	R	2161.0229	2143.0125	2143.9958	1081.0115	18
6	823.4097	805.4062	806.3812	412.2084	E	2004.9219	1986.9114	1987.8949	1002.9609	17
7	938.4370	920.4266	921.4072	469.7186	N(+.98)	1875.8793	1857.8688	1858.8523	938.4370	16
8	995.4666	977.4541	978.4316	498.2354	G	1760.8523	1742.8418	1743.8253	880.9261	15
9	1137.5692	1119.5587	1120.5422	569.2846	K(+14.02)	1703.8308	1685.8203	1686.8038	852.4154	14
10	1297.5974	1279.5966	1280.5729	649.3026	C(+57.02)	1561.7202	1543.7097	1544.6932	781.3601	13
11	1410.6787	1392.6735	1393.6570	705.8420	L	1401.6896	1383.6791	1384.6626	701.3448	12
12	1538.7417	1520.7321	1521.7156	769.8713	Q	1288.6055	1270.5950	1271.5779	644.8027	11
13	1685.8109	1667.8004	1668.7839	843.4091	F	1160.5415	1142.5349	1143.5200	580.7735	10
14	1772.8390	1754.8325	1755.8160	886.9215	S	1013.4783	995.4666	996.4515	507.2393	9
15	1885.9270	1867.9165	1868.9000	943.4635	L	926.4466	908.4360	909.4196	463.7233	8
16	1986.9747	1968.9642	1969.9478	993.9874	T	813.3617	795.3518	796.3354	407.1812	7
17	2088.0225	2070.0120	2070.9954	1044.5112	T	712.3150	694.3053	695.2878	356.6574	6
18	2251.0857	2233.0752	2234.0586	1126.0428	Y	611.2671	593.2563	594.2401	306.1335	5
19	2379.1807	2361.1702	2362.1536	1190.0903	K	448.2038	430.1936	431.1768	224.6019	4
20	2436.2021	2418.1917	2419.1750	1218.6011	G	320.1088	302.0983	303.0818	160.5544	3
21	2565.2446	2547.2341	2548.2175	1283.1223	E	263.0873	245.0770	246.0603	132.0437	2
22					D	134.0448	116.0342	117.0178	67.5224	1



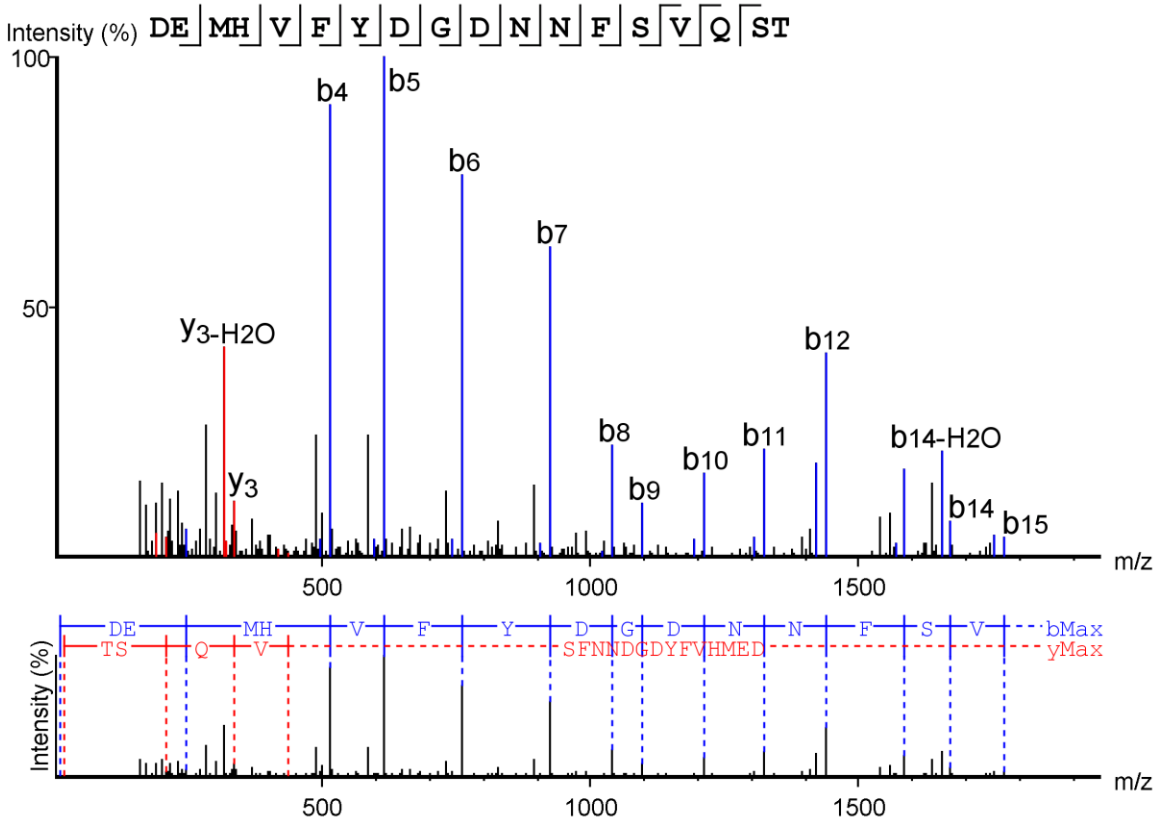
a3



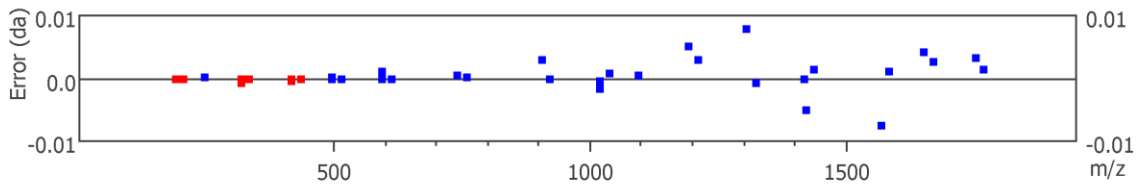
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					7
2	245.0769	227.0668	228.0503	123.0387	E	825.3600	807.3494	808.3330	413.1800	6
3	376.1174	358.1071	359.0909	188.5589	M	696.3177	678.3068	679.2904	348.6587	5
4	513.1761	495.1656	496.1498	257.0884	H	565.2772	547.2666	548.2499	283.1384	4
5	612.2446	594.2343	595.2184	306.6226	V	428.2189	410.2074	411.1910	214.6090	3
6	759.3129	741.3013	742.2856	380.1568	F	329.1496	311.1390	312.1226	165.0748	2
7					Y	182.0814	164.0706	165.0546	91.5406	1



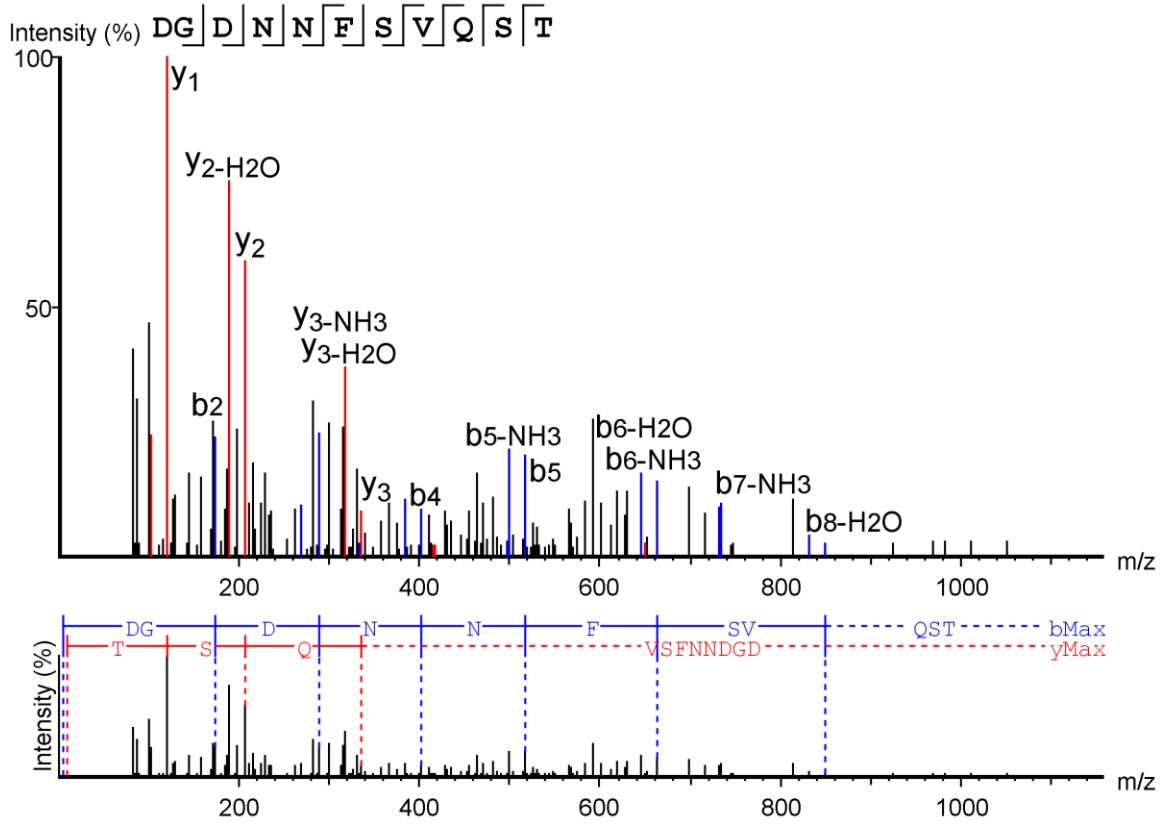
a4



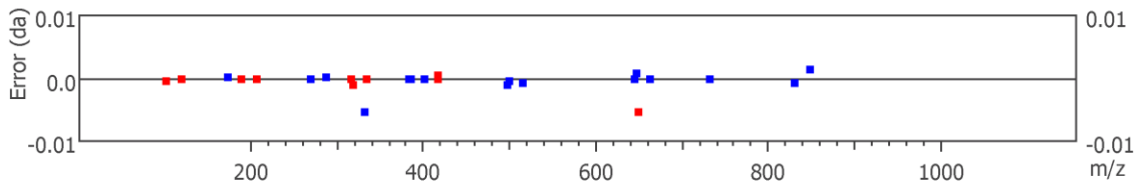
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					18
2	245.0770	227.0668	228.0503	123.0387	E	1989.8282	1971.8177	1972.8013	995.4141	17
3	376.1178	358.1073	359.0909	188.5589	M	1860.7858	1842.7753	1843.7588	930.8929	16
4	513.1765	495.1659	496.1494	257.0884	H	1729.7452	1711.7347	1712.7183	865.3726	15
5	612.2449	594.2344	595.2167	306.6226	V	1592.6863	1574.6758	1575.6593	796.8431	14
6	759.3130	741.3021	742.2866	380.1568	F	1493.6179	1475.6074	1476.5909	747.3090	13
7	922.3766	904.3630	905.3499	461.6884	Y	1346.5494	1328.5389	1329.5225	673.7747	12
8	1037.4028	1019.3937	1020.3784	519.2019	D	1183.4862	1165.4757	1166.4592	592.2431	11
9	1094.4244	1076.4148	1077.3983	547.7126	G	1068.4592	1050.4487	1051.4323	534.7296	10
10	1209.4490	1191.4363	1192.4253	605.2261	D	1011.4377	993.4272	994.4108	506.2189	9
11	1323.4958	1305.4766	1306.4681	662.2476	N	896.4108	878.4003	879.3839	448.7054	8
12	1437.5363	1419.5275	1420.5162	719.2690	N	782.3679	764.3574	765.3409	391.6840	7
13	1584.6051	1566.5959	1567.5872	792.8032	F	668.3250	650.3144	651.2980	334.6625	6
14	1671.6357	1653.6237	1654.6116	836.3193	S	521.2566	503.2460	504.2296	261.1283	5
15	1770.7053	1752.6930	1753.6799	885.8535	V	434.2243	416.2141	417.1980	217.6123	4
16	1898.7655	1880.7550	1881.7385	949.8828	Q	335.1561	317.1456	318.1299	168.0781	3
17	1985.7976	1967.7871	1968.7706	993.3988	S	207.0977	189.0870	190.0705	104.0488	2
18					T	120.0655	102.0550	103.0385	60.5328	1



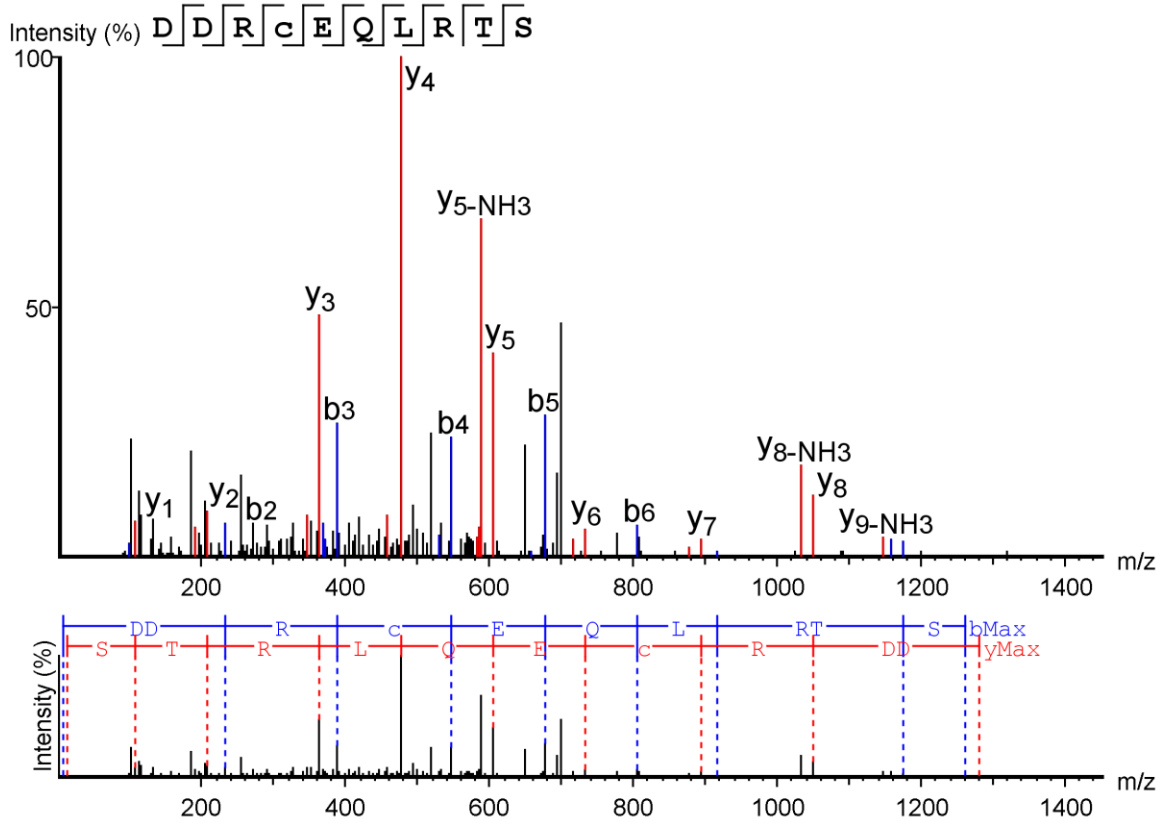
a5



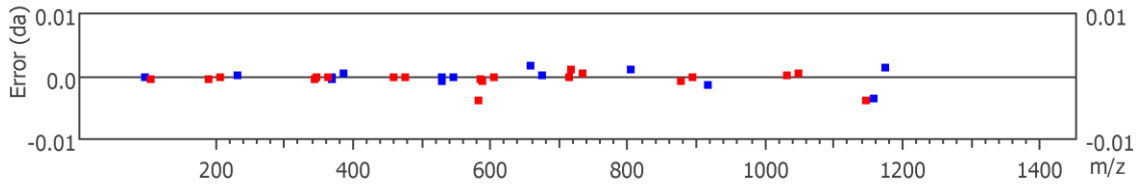
#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0242	99.0078	58.5174	D					11
2	173.0558	155.0457	156.0292	87.0281	G	1068.4592	1050.4487	1051.4323	534.7296	10
3	288.0827	270.0726	271.0562	144.5416	D	1011.4377	993.4272	994.4108	506.2189	9
4	402.1262	384.1157	385.0992	201.5630	N	896.4108	878.4003	879.3839	448.7054	8
5	516.1697	498.1595	499.1425	258.5845	N	782.3679	764.3574	765.3409	391.6840	7
6	663.2377	645.2269	646.2094	332.1241	F	668.3250	650.3197	651.2980	334.6625	6
7	750.2695	732.2591	733.2424	375.6347	S	521.2566	503.2460	504.2296	261.1283	5
8	849.3361	831.3281	832.3109	425.1689	V	434.2245	416.2131	417.1977	217.6123	4
9	977.3964	959.3859	960.3694	489.1982	Q	335.1562	317.1457	318.1303	168.0781	3
10	1064.4285	1046.4180	1047.4015	532.7142	S	207.0977	189.0872	190.0705	104.0488	2
11					T	120.0657	102.0553	103.0385	60.5328	1



a6



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	116.0348	98.0243	99.0078	58.5174	D					10
2	231.0613	213.0511	214.0347	116.0309	D	1164.5426	1146.5321	1147.5195	582.7752	9
3	387.1621	369.1521	370.1361	194.0814	R	1049.5149	1031.5051	1032.4880	525.2578	8
4	547.1935	529.1831	530.1672	274.0967	C(+57.02)	893.4144	875.4040	876.3884	447.2073	7
5	676.2355	658.2235	659.2090	338.6180	E	733.3831	715.3734	716.3556	367.1919	6
6	804.2933	786.2841	787.2676	402.6473	Q	604.3412	586.3311	587.3150	302.6707	5
7	917.3801	899.3681	900.3517	459.1893	L	476.2828	458.2724	459.2557	238.6414	4
8	1073.4797	1055.4692	1056.4528	537.2399	R	363.1986	345.1885	346.1720	182.0993	3
9	1174.5259	1156.5170	1157.5039	587.7637	T	207.0978	189.0874	190.0705	104.0488	2
10					S	106.0503	88.0393	89.0229	53.5249	1



5.6.5 *Fragment ion spectra for determination of leucine and isoleucine residues in peptides with two ambiguous leucine or isoleucine residues.*

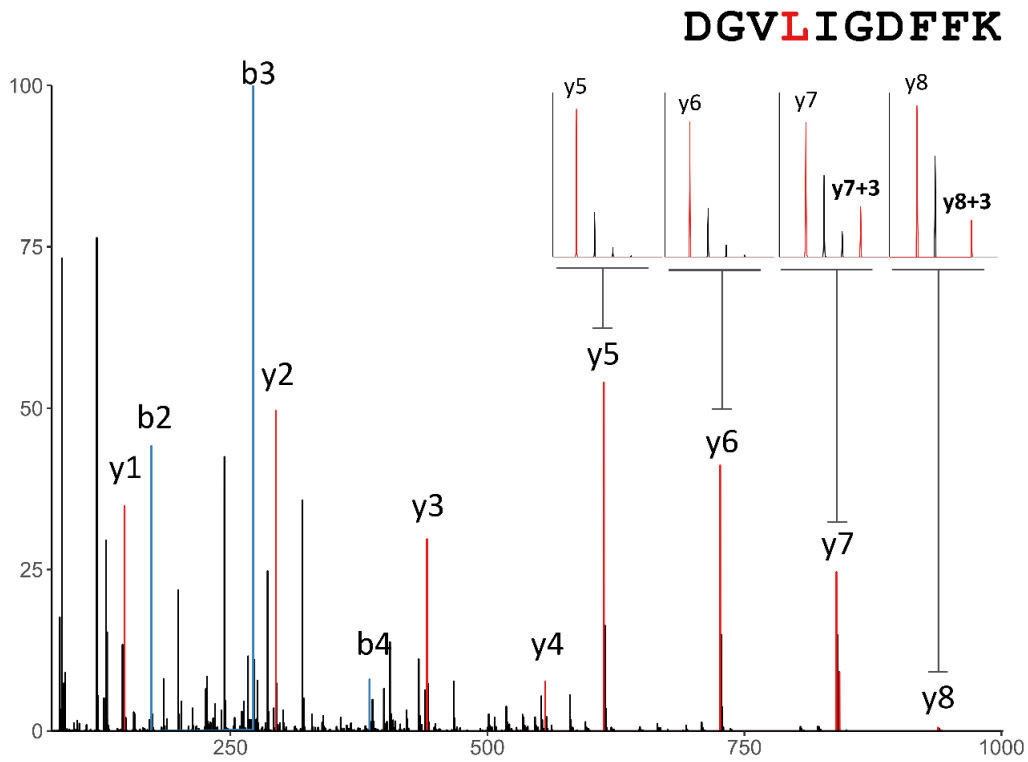


Figure 5.21 Fragment ion spectrum of the peptide DGVLIGDFFK. Inset are enlargements of isotope distributions of two y ions without the incorporation of a heavy leucine (y5 and y6) and two y ions after incorporation of heavy leucine (y7 and y8), both of which show an increase in intensity of +3 Da.

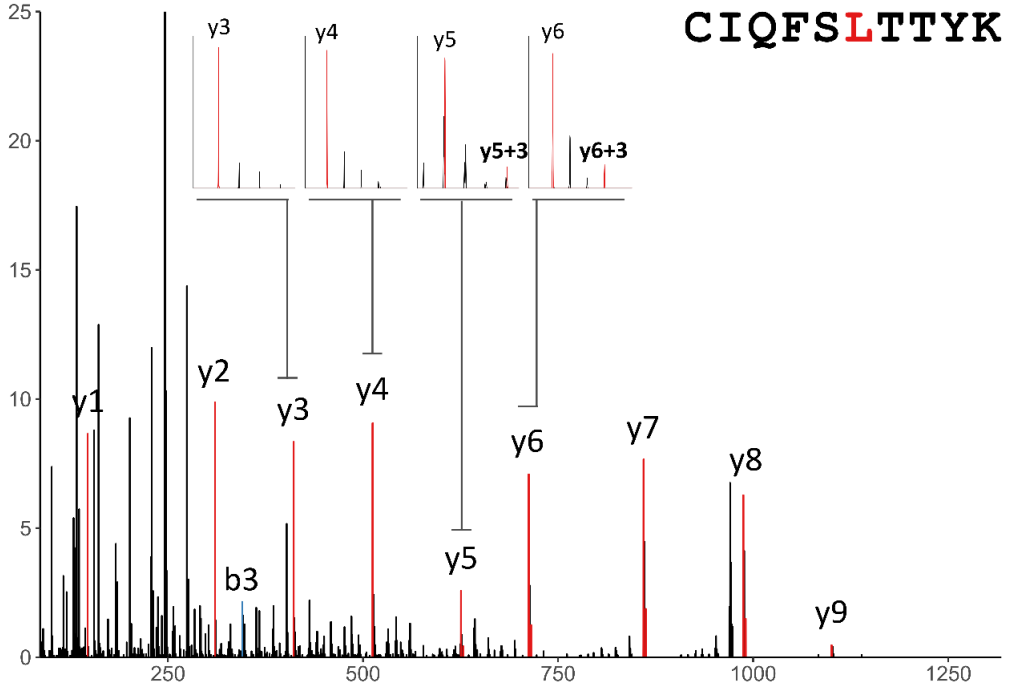


Figure 5.22 Fragment ion spectrum of the peptide CIQFSLTTYK. Inset are enlargements of isotope distributions of two y ions without the incorporation of a heavy leucine (y3 and y4) and two y ions after incorporation of heavy leucine (y5 and y6), both of which show an increase in intensity of +3 Da.

5.6.6 Table of accession numbers for multiple sequence alignment

Table 5.2 List of protein accessions used for multiple sequence alignment and subsequent phylogenetic analysis of lipocalins with known or suspected chemosignalling functions. Representatives from each sub-family were included, from species in which the protein sequence has been manually annotated and reviewed, from the SwissProt database. Also included were lipocalins homologous to vulpeculin identified in the unannotated genomes of the koala (*P. cinereus*), Tasmanian devil (*S. harrisii*) and short-tailed opossum (*M. domestica*) by BLAST search of the NCBI database.

Species	Common name	MUPs	OBPs & OBP-like	Allergen/Salivary Lipocalins	Epididymal-specific lipocalins	Trichosurin
Homo sapiens	Human		Q9NY56 2a Q9NPH6 2b		Q8WX39	
Mus musculus	House mouse	MUPs 1-21	Q9D3H2 1a A2AEP0 1b Q8K1H9 2a A2BHR0 2b O08976 (Probasin)		Q9D267	
Rattus norvegicus	Brown rat	Q8K1Q6 P02761 Q63025 F7FIH7 Q9JJI3 Q8K1Q6 P02761 Q63024 Q63025	P08937 Q63613 B3EY84 P15399 (Probasin)		B3EY87	

		F7FIH7 Q9JJI3 Q78E14 Q63213 F1M6Y6 F8WFF8 Q9JJI1 Q9JJI2 Q9JJI0 Q9JJH9				
Felis catus	Domestic cat			Q5VFH6	M3WIX8	
Canis lupus familiaris	Dog		J9P950	H2B3G5	F6X9G7	
Sus scrofa	Wild boar		P81245	P81608	F1S003	
Ovis aries	Sheep		W5PHS2	W5P8Y1	W5P4T6	
Bos taurus	Cow		P07435	E1BJP1	E1BLF9	
Equus callabus	Horse		F7DPI0	Q95182	F7AXE4	
Cavia porcellus	Guinea pig		H0W3W3	S0BDX9	H0VW61	
Mesocricetus auratus	Golden hamster		Q9Z1I7 (Aphrodisin) Q99MG7 (FLP) Q9QXU1 (MSP)	A0A0H4SQI0 T2B8F6	A0A1U8CGZ8	
Cricetus cricetus	European hamster				P09465	

Cricetulus griseus	Chinese hamster			G3HPK8	G3ID51	G3IJ88
Phodopus roborovskii	Roborovski hamster		Roborovskin			
Myodes glareolus	Bank vole		Glareosin D3VW64 D3VW63 D3VW62			
Elephas maximus	Asian elephant		Q8HZN2			
Oryctolagus cuniculus	European rabbit			U6C8D6		
Mustela putorius furo	Ferret			M3Y8P3		
Trichosurus vulpecula	Common brushtail possum					Q29147
Monodelphis domestica	Gray short-tailed opossum	*From NCBI:* XP_007475416.1 XP_007475409.1 XP_007475414.1 XP_007475412.1 XP_007475429.1		*From UniProt* K7E3M5 K7E641 F7F0X2 F6VSN8 F7F1B2 F7F0W8		
Sarcophilus harrisii	Tasmanian devil	*From NCBI:*		*From UniProt*		

		XP_012396095.1 XP_012396096.1 XP_012409487.1 XP_012396091.1 XP_012396094.1 XP_003758835.1 XP_023350144.1	G3W2D8 G3VNH1 G3VM27 G3VPW6 G3VEI0
Phascolarctos cinereus	Koala	*From NCBI:* XP_020837510.1 XP_020837036.1 XP_020837035.1 XP_020837051.1 XP_020837067.1 XP_020837033.1 XP_020837509.1	

5.6.7 *Multiple sequence alignment of lipocalins*

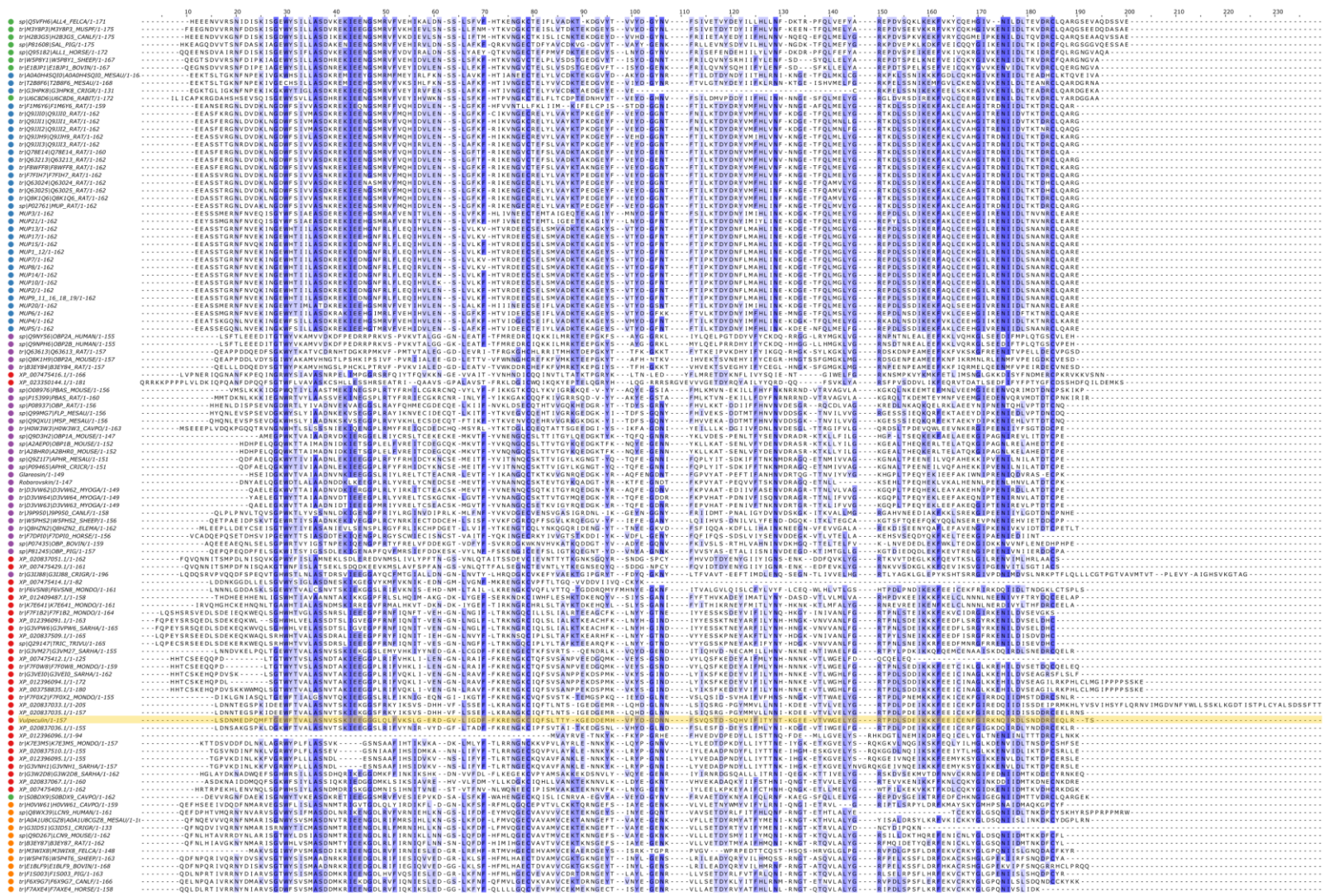


Figure 5.23 Multiple sequence alignment of lipocalin proteins from the table above, for use in phylogenetic analysis.

6 Characterisation of urinary WFDC12 in nocturnal basal primates, mouse lemurs (*Microcebus spp.*)

www.nature.com/scientificreports

OPEN

Characterisation of urinary WFDC12 in small nocturnal basal primates, mouse lemurs (*Microcebus spp.*)

Received: 13 September 2016
Accepted: 17 January 2017
Published: 22 February 2017

Jennifer Unsworth[1,*], Grace M. Loxley[1,*], Amanda Davidson[2], Jane L. Hurst[2], Guadalupe Gómez-Baena[1], Nicholas I. Mundy[3], Robert J. Beynon[1], Elke Zimmermann[4] & Ute Radespiel[4]

Mouse lemurs are basal primates that rely on chemo- and acoustic signalling for social interactions in their dispersed social systems. We examined the urinary protein content of two mouse lemurs species, within and outside the breeding season, to assess candidates used in species discrimination, reproductive or competitive communication. Urine from *Microcebus murinus* and *Microcebus lehilahytsara* contain a predominant 10 kDa protein, expressed in both species by some, but not all, males during the breeding season, but at very low levels by females. Mass spectrometry of the intact proteins confirmed the protein mass and revealed a 30 Da mass difference between proteins from the two species. Tandem mass spectrometry after digestion with three proteases and sequencing *de novo* defined the complete protein sequence and located an Ala/Thr difference between the two species that explained the 30 Da mass difference. The protein (mature form: 87 amino acids) is an atypical member of the whey acidic protein family (WFDC12). Seasonal excretion of this protein, species difference and male-specific expression during the breeding season suggest that it may have a function in intra- and/or intersexual chemical signalling in the context of reproduction, and could be a cue for sexual selection and species recognition.

Chemical signalling is the evolutionarily oldest and most widespread communication mode among animals[1]. Chemical signals are highly variable, ranging from small volatile molecules to large non-volatile proteins. They are perceived by two olfactory systems; the main olfactory epithelium in the nasal cavity and the vomeronasal organ that is specialized in pheromone detection in mammals[2]. Chemical signals are typically deposited by specific glands or *via* urine or other body fluids and faeces. However, whereas there is a large body of literature on species-specific olfactory behaviours (e.g. scent marking), there is typically much less knowledge about the information content or biochemical composition of semiochemicals. A notable exception is provided by the domestic house mouse, the first mammal for which there is a comprehensive understanding of the role of low molecular weight volatile components, proteins and their interplay in semiochemistry. In the house mouse, proteins act both as pheromone binders and also as pheromones in their own right[3–6]. However, the extent to which such protein-mediated semiochemistry operates across the animal kingdom outside the myomorph rodents is unknown. The proteins of mouse urine, the major urinary proteins (MUPs), have multiple roles, including individual recognition, inter-male aggression, kin recognition and induction of learning[3,4,7–12]. MUPs are highly polymorphic[13–15], eliciting the combinatorial complexity that is needed for such subtle semiochemical effects[15–18] Hurst *et al.*, in revision.

Mouse lemurs (*Microcebus spp.*) are small nocturnal solitary foragers constituting one primate genus endemic to Madagascar. They strongly rely on olfactory and acoustic signalling in adaptation to their dispersed nocturnal life style[19,20]. Scent marking behaviour of mouse lemurs is rich and includes urine washing, anogenital marking and substrate rubbing[20,21]. Mouse lemurs possess a functional vomeronasal organ (VNO[22,23]) and have a repertoire of over 200 vomeronasal receptor genes[24,25] that are expressed in the VNO but also partially in the main

[1]Centre for Proteome Research, Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK. [2]Mammalian Behaviour & Evolution Group, Institute of Integrative Biology, University of Liverpool, Leahurst Campus, Neston, CH64 7TE, UK. [3]Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK. [4]Institute of Zoology, University of Veterinary Medicine, Hannover, Buenteweg 17, 30559 Hannover, Germany. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to R.J.B. (email: r.beynon@liverpool.ac.uk)

olfactory epithelium[22]. Olfaction may be involved in various contexts of intraspecific communication, as in marking group ownership of sleeping sites[19], or in many aspects of reproduction, and may be used within and between the sexes[20,26–29].

Mouse lemurs reproduce strictly seasonally and produce one or two litters per year that are reared during the resource-rich rainy season[30–32]. Their mating system is a multi-male/multi-female system that involves male mouse lemurs competing intensely for access to oestrous females during the reproductive season[26,33,34]. On the other hand, females exert some level of mate choice[35–37] while having to assure fertilization during the one or two very short receptive periods per year[28]. Physiological changes prior to the breeding season include decreased body mass (due to loss of adipose tissue) and increased testis size[38,39]. This testis size increase can be reversed subsequently in subordinate males by exposure to urine from a dominant male[40].

A preliminary survey of *Mup* genes across non-rodent mammals proposed that two genes encoding MUPs could be located within the first draft of the grey mouse lemur (*Microcebus murinus*) genome[13]. However, there are no data on expression and tissue specificity for these putative proteins. In this study, we examined the urine of two species of mouse lemur – *Microcebus murinus* and *Microcebus lehilahytsara* – with the aim of identifying urinary proteins that have the potential to be involved in chemosignalling in a reproductive context or in species recognition. We focussed on two allopatrically-distributed species that are maintained under standardized conditions in the laboratory to identify potential species-specific differences. We report that mouse lemur urine does not contain detectable levels of MUPs, notwithstanding the two *Mup* gene paralogues reported by Logan and colleagues[13]. However, the urine of males from both species sometimes contains high levels of a low molecular weight, seasonally expressed protein. We gained sufficient protein-level information to permit identification of this protein and speculate on likely functions.

Materials and Methods

Animals. Urine samples were obtained from 24 (14 male, 10 female) *M. murinus* and from 13 (8 male, 5 female) *M. lehilahytsara* that belong to the captive mouse lemur colony at the Institute of Zoology, University of Veterinary Medicine Hannover. Samples were taken during the reproductive period (March–May 2011 or March–May 2012) and during the non-reproductive period (September 2011–January 2012). Housing conditions[41,42] were in accordance with European Directive 2010/63/EU on the protection of animals used for scientific purposes, the German Animal Welfare Act, and corresponding section for animals used for scientific purposes and licensed by the Bezirksregierung Hannover (reference number: AZ 33.9-42502-05-10A080) and by the Ordnungsamt, Gewerbe- und Veterinärabteilung, Landeshauptstadt Hannover (AZ 42500/1H). Animals were housed in groups of two or three individuals of the same or opposite sex (depending on the breeding management decisions). The non-invasive procedure was approved by the animal welfare officer of the University of Veterinary Medicine Hannover, Foundation as well as by the State of Lower Saxony Office for Consumer Protection and Food Safety (LAVES; approval date: April 28, 2014; number: 33.12-42502-04-14/1454), which is the responsible agency of the State Lower Saxony for approval of animal studies according to the German Animal Welfare Act (TierSchG). The study is in accordance with the recommendations of the Weatherall report, "*The use of non-human primates in research*" (https://www.mrc.ac.uk/documents/pdf/the-use-of-non-human-primates-in-research).

The animals are kept in same or mixed-sex pairs or mixed-sex pairs with offspring. They are housed in cages measuring 150 × 62.5 × 80 cm for a single individual or 150 × 150 × 80 cm and 140 × 170 × 70 cm, respectively, for a pair. The diet consists of seasonal fresh fruits and vegetables, dried fruits, and mealworms every other day. On non-mealworm days, the animals are offered milk porridge enriched with vitamins, minerals, and albumin[41]. Water is available *ad libitum*. Environmental enrichment is provided within the housing cages, branches and hollow cylinders allow the animals to climb and hide. To simulate the natural condition, where mouse lemurs sleep, rest and rear their offspring in tree holes[43], each cage is equipped with several sleeping boxes (20 × 11 × 11 cm each) that are also occasionally used to collect urine samples. Urine sampling also took place opportunistically by pipetting during the weekly handling routine handling when each individual is constrained for one to two minutes in hand when body mass, health condition and reproductive status are assessed. Suffering can be stated to be minimal, since the animals are used to these weekly routines.

Urine collection. Samples were obtained by direct sampling during weekly handling routines or with a urine collection box. During the handling routines, the urine was collected with a disposable 1 mL pipette whenever the animals urinated spontaneously. Alternatively, the animals were confined to a urine collection box (20.5 × 12 × 13 cm, length × depth × height) for 30–90 min at the beginning of their activity period. The urine passed through a mesh at the bottom of the box and was collected by a stainless steel funnel. The collection box and funnel were cleaned thoroughly after every use. All urine samples were frozen at −20 °C within 2 h of collection. For some lemurs, multiple samples were collected (see Supplementary Data) but have been treated as independent measures.

Protein characterization. Total protein concentration was measured using a Coomassie Plus protein assay kit (Pierce, Rockford, USA) using bovine serum albumin as standard. SDS-PAGE was conducted using standard protocols. Creatinine concentration was measured by the alkaline picrate assay kit from Sigma-Aldrich.

Electrospray ionisation mass spectrometry. Urine samples were diluted in formic acid (0.1% (v/v) in HPLC grade water) to a protein concentration of approximately 5 pmol/µl. The samples were injected onto a C4 desalting trap (Waters MassPREP™ Micro desalting column, 2.1 × 5 mm, 20 µm particle size, 1000 Å pore size) (Waters, Manchester, UK) that was fitted on a Waters nano ACQUITY Ultra Performance liquid chromatography® (UPLC®) system. The chromatography system was coupled to a Waters SYNAPT™ G1 QTof mass spectrometer fitted with an electrospray source. Protein was eluted over a 10 min acetonitrile (ACN) gradient

(5–95% (v/v)) at 40 µl/min. Data were collected between 500–3500 m/z. The data were processed using maximum entropy deconvolution (MAX ENT 1, Mass Lynx version 4.1, Waters) at 0.5 Da/channel over a mass range of 8500–10000 Da. The mass spectrometer was calibrated externally with horse heart myoglobin (1 pmol/µl, Sigma).

In-gel digestion. Pieces of SDS-PAGE gel containing protein bands were washed repeatedly with 25 mM NH_4HCO_3, ACN (50:50) for 15 min at 37 °C until the gel pieces were fully destained. The gel plugs were then reduced in dithiothreitol (10 mM) at 60 °C for 1 h. The dithiothreitol solution was discarded and cysteine residues were alkylated with iodoacetamide (25 µl, 55 mM) in the dark at room temperature for 45 min. The recovered gel pieces were dehydrated in ACN for 15 min at 37 °C. Proteolytic enzymes – trypsin, endoproteinase LysC or endoproteinase GluC (each 10 µl, 10 ng/ml) were added to each of the gel pieces and incubated for 16 h. The digestion was stopped by the addition of formic acid (final concentration 1% v/v).

MALDI-TOF analysis. The peptide mixtures from in-gel digestion were analysed by MALDI-TOF-MS on a Bruker ultrafleXtreme™ mass spectrometer in reflectron mode. Samples were mixed with MALDI matrix (saturated solution of α-cyano-4-hydroxycinnamic acid in 50% (v/v) ACN/0.2%(v/v) trifluoroacetic acid (TFA)) in a 1:1 ratio and spotted onto a target plate before being left to air dry. The laser frequency was 1000 Hz, laser energy 27% of maximum and 500 laser shots were collected per spectrum, between 800–4000 m/z.

In solution digestion. Samples, diluted in 25 mM NH_4HCO_3 to 10 µg protein in a 50 µl digest were incubated with RapiGest™ SF Surfactant (0.05% w/v final concentration, Waters, Manchester, UK) at 80 °C for 10 min. The samples were then reduced with dithiothreitol (3 mM final concentration) at 60 °C for 10 min followed by alkylation with iodoacetamide (9 mM final concentration) in the dark at room temperature for 30 min. The protease, either trypsin (0.2 µg/µl diluted in 25 mM NH_4HCO_3) or endoproteinase LysC (0.1 µg/µl diluted in 25 mM Tris HCl pH 8.5), was added to the digests at a substrate:enzyme ratio of 50:1 and left to incubate for 16 hours. Following incubation, a small portion of the digested material was removed to run on an SDS-PAGE gel to check for complete digestion. The rest of the digest was treated with TFA (to a final concentration of 0.5% v/v) and incubated at 37 °C for 45 min to precipitate the RapiGest™ SF Surfactant prior to LC-MS analysis. The samples were then centrifuged at 10,000 rpm for 15 min and the supernatant transferred to a fresh 0.5 ml Eppendorf.

LC-MS analysis. Digests were analysed using an Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific, Hemel Hempstead, UK) coupled to a QExactive mass spectrometer (Thermo Fisher Scientific, Hemel Hempstead, UK). Peptides were loaded onto a trap column (Acclaim PepMap 100, 2 cm × 75 m inner diameter, C18, 3 µm particle size, 100 Å pore) at 5 µL/min with an aqueous solution containing 0.1%(v/v) TFA and 2%(v/v) ACN. After 3 min, the trap column was set in-line with an analytical column (Easy-Spray PepMap® RSLC 15 cm × 75 µm inner diameter, C18, 2 µm, 100 Å) (Dionex/Thermo Fisher). Peptides were eluted by using an appropriate mixture of solvents A and B. Solvent A was 0.1%(v/v) formic acid in HPLC-grade water, and solvent B was 0.1% (v/v) formic acid in HPLC grade acetonitrile 80% (v/v). Separations were performed by applying a linear gradient of 3.8% to 50% solvent B over 35 min at 300 nL/min followed by a washing step (5 min at 99% solvent B) and an equilibration step (15 min at 3.8% solvent B). The mass spectrometer was operated in data dependent positive (ESI+) mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (300–2000 m/z) were acquired in the Orbitrap at 70,000 resolution (200 m/z) after accumulation of ions to 1×10^6 target value based on predictive automatic gain control (AGC) values from the previous full scan. Dynamic exclusion was set to 20 s. The 10 most intense multiply charged ions ($z \geq 2$) were sequentially isolated and fragmented in the octopole collision cell by high energy collisional dissociation (HCD) with a fixed injection time of 120 ms and detected in the Orbitrap at 35,000 resolution. The mass spectrometer conditions were as follows: spray voltage, 1.9 kV, no sheath or auxillary gas flow; heated capillary temperature, 250 °C; normalised HCD collision energy 30%. The MS/MS ion selection threshold was set to 1×10^4 counts and a 2 m/z isolation width was set. Peak lists were generated using Proteome Discoverer (Thermo Fisher) and MASCOT (Matrix Science, London) as search engine. The search parameters were set to accept 1 missed cleavage, a fixed modification of carbamidomethylation of cysteine and variable methionine oxidation. Precursor and fragment ion error tolerances were set to 10 ppm and 0.01 Da respectively. Fragmentation was set to HCD. The MS/MS data were first searched against a database of all mammalian protein sequences in the SwissProt database, prior to searching all primate protein sequences from UniProt. Once the sequence of the protein was determined, a database entry was created containing the protein sequence from both mouse lemur species to assist in peptide coverage analysis.

Data analysis. Urinary protein output (µg protein/µg creatinine) was log transformed to meet assumptions of parametric analysis (Shapiro-Wilks, $P > 0.05$). To take into account the variable number of samples available from different individuals, we used R (version 3.2.5) and the *lme4* test package[44] to examine the effects of donor sex and season on protein output in a linear mixed effects analysis, with individual donor and species included as random effects (insufficient breeding season samples were available from *M. lehilahytsara* to include species as a main effect in analysis). As there was a nearly significant interaction between donor sex and breeding season, separate models examined the effect of breeding season on protein output within each sex. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity and the distribution of residuals from each model approximated normality (Shapiro-Wilks, $P > 0.05$). Likelihood ratio tests compared the full model against a reduced model without the effect of interest using the anova function.

Sequencing analysis *de novo* of the mouse lemur protein was assisted by PEAKS software[45] for proteomics (Bioinformatics Solutions Inc, Canada). Precursor and fragment ion error tolerances were set to 10 ppm and 0.01 Da respectively. Post translational modifications, carbamidomethylation (fixed modification) and oxidation of methionine (variable modification) residues were also included in the processing set-up. Fragmentation type was set to higher-energy C-trap dissociation (HCD). The average local confidence score – a score assigned by

PEAKS which reflects the likelihood of a peptide sequence being correct – was set to a 55% cut off as recommended by the software vendor. The *de novo* sequenced peptides were then analysed using BLAST, (http://blast.ncbi.nlm.nih.gov), and compared to annotated exon/intron structure of *WFDC12* entries in Ensembl. The complete mature protein sequence was used to direct homology modelling using the Phyre2 server[46]. To confirm the sequence, the LC-MS/MS data was then searched against the newly constructed protein sequence in PEAKS, using database search tools PEAKS DB, which matched the *de novo* sequences generated against a database of all primate protein sequences in UniProt with the addition of the constructed protein sequence, and SPIDER, to search for mutations and sequencing errors. Search parameters included precursor and fragment ion error tolerances of 10 ppm and 0.01 Da, respectively. For the PEAKS DB search, a fixed modification of carbamidomethylation was set, and a maximum of 3 variable modifications (restricted to oxidation of methionine) per peptide was allowed. No non-specific cleavage was allowed, and the maximum number of missed cleavages was set to 2. The same error tolerances were used for the PEAKS PTM search[47], which matched additional peptides from the *de novo* sequences to the database by searching for mass changes corresponding to a default list of 484 variable modifications, and the SPIDER search[48], which attempts to find mutations and sequencing errors based on combining the information gained from the *de novo* sequences generated and the database sequence. For both searches, a threshold peptide confidence score (−10 logP) of 30 was used. PEAKS PTM also maintained the cleavage parameters set from the PEAKS DB search, whereas SPIDER automatically searched for non-specific cleavages. A false discovery rate of 1% was used for each search tool, and an ALC (average local confidence) cutoff score of 50% was used for *de novo*-only peptides. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[49] partner repository with the dataset identifier PXD005196 and 10.6019/PXD005196.

Protein modelling was conducted using the Phyre2 server (http://www.sbg.bio.ic.ac.uk/~phyre2) in intensive mode[46].

Results and Discussion

Urine samples (116 samples in total, from 37 individuals) were collected from both male and female mouse lemurs (*M. murinus and M. lehilahytsara* data are combined in Fig. 1) during the breeding and the non-breeding season (details of samples are in Supplementary Table S1). The urine was initially analysed by 1D SDS-PAGE (Fig. 1a). No band corresponding to a MUP-like protein at 19 kDa was observed, but there was a predominant protein band at approximately 10 kDa present in some male samples from either species, exclusively during the reproductive season. No predominant protein was identified in female mouse lemur urine from either species, whether in or outside of the reproductive season (Fig. 1a). The concentration of protein varied between 0.08 mg/ml and 2.2 mg/ml across all animals and samples. However, there was a strong relationship between protein and creatinine concentration ($r^2 = 0.66$, [protein] = 0.26 * [creatinine] + 40.6; Fig. 1b), most likely reflecting differences in urine dilution between samples. To correct for urine dilution, protein output was expressed as μg protein/μg creatinine. The effects of donor sex and breeding season on urinary protein output (data log transformed to meet assumptions for parametric analysis) were assessed using linear mixed effects models, with individual donor and species included as random effects. Confirming our initial observations from SDS-PAGE bands, seasonal effects on protein output appeared to vary between the sexes (interaction between breeding season and sex: $^2 = 3.00$, 1df, $P = 0.08$; Fig. 1c). Separate models were thus run to assess seasonal variation in urinary protein output within each sex. Among males, protein output was significantly elevated during the breeding season compared to the non-breeding season, a difference that was highly significant (effect of season: $^2 = 12.99$, 1df, $P = 0.0003$, Fig. 1c). In the breeding season, male output ranged from 0.13 to 5.46 mg protein/mg creatinine (mean ± SEM: 1.19 ± 0.34, n = 18 samples from 10 individuals), compared to a non-breeding season range of 0.10 to 1.12 (mean ± SEM: 0.43 ± 0.04, n = 50 samples from 22 individuals). By contrast, protein output showed no seasonal variation among females (effect of season: $^2 = 1.12$, 1df, $P = 0.29$, Fig. 1c) and was very similar to the output of non-breeding males (female mean ± SEM: 0.41 ± 0.05, n = 48 samples from 15 individuals, range 0.09 to 1.76). Among ten males that were sampled during both the breeding and non-breeding season, eight had higher urinary protein output in breeding season samples (Fig. 1b,d). Breeding season samples from these males revealed a strong 10 kDa band on SDS-PAGE. Of eight breeding season *M. murinus* males, three had exceptionally high levels of protein in urine, as did both of the breeding season *M. lehilahytsara* males that were sampled.

We have previously used mass spectrometry-based intact mass analysis to successfully characterise predominant urinary proteins in other species, and to ascertain the extent of any mass-resolvable heterogeneity[50–52]. Male mouse lemur urinary proteins were therefore analysed by electrospray ionisation mass spectrometry of diluted and desalted urine (Fig. 2). A protein of molecular mass 9,388 Da was identified in all male *M. murinus* samples that also demonstrated a strong band at approx. 10 kDa on SDS-PAGE – in turn these were all from males in the breeding season. A second mass peak, 16 Da heavier, was also identified; we surmised that this was probably due to partial oxidation of the main protein peak. Analysis of the urinary proteins in *M. lehilahytsara* urine similarly revealed a dominant protein of molecular mass 9,418 Da (30 Da heavier than the protein from *M. murinus*) in the male-derived samples that also expressed a 10 kDa SDS-PAGE band. These too were all from breeding season males. This protein mass was also partnered by a second mass peak, 16 Da heavier, as observed with *M. murinus*. In both species, the mass profile was simple, exhibiting none of the heterogeneity that we have previously observed for the MUPs of the domestic mouse[11,16,52].

To investigate the similarity of the 10 kDa protein in the two species, urine samples from breeding season males were resolved by SDS-PAGE, proteolysed by in-gel digestion and the tryptic peptides were analysed by MALDI-TOF mass spectrometry (Fig. 3). Many of the peptides were of the same mass in the digest from either species, very good evidence for the proteins being orthologues. However, two peptides (singly charged ions, $[M+H]^+$, at m/z 991 and m/z 1536) in the peptide mass fingerprints from *M. lehilahytsara* samples were 30 Da

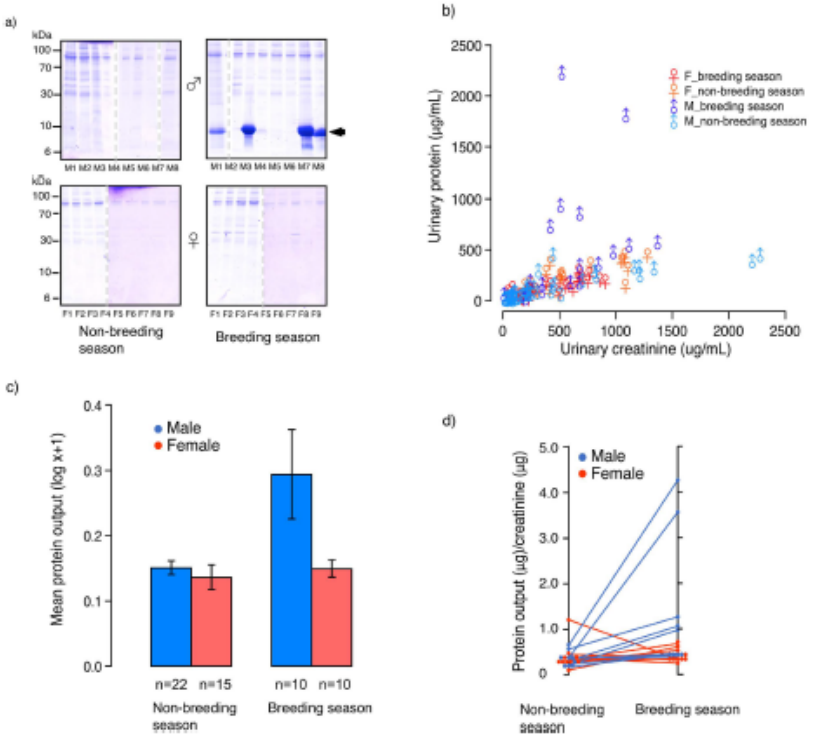


Figure 1. Protein content of mouse lemur urine. A group of 116 male and female mouse lemur (*M. murinus* and *M. lehilahytsara*) urine samples were analysed by SDS-PAGE (panel a, samples from several individuals, males: M1 to M8, females F1 to F9). Total urinary protein concentration generally increased with urinary creatinine, which provides a measure of urine dilution (panel b, symbols and colours represent donor sex and season). Some males in the breeding season had much higher levels of urinary protein than explained by urine dilution. The urinary protein output (expressed as µg protein/µg creatinine, averaged over replicate samples from the same individual and season) was higher in the breeding than in the non-breeding season among males, while there was no seasonal effect on female output (panel c, data are means ± sem, n indicates total number of individuals sampled). Multiple samples obtained from some individuals (10 males, 10 females) allowed urinary protein output to be compared between breeding and non-breeding season within the same individual (panel d).

heavier than cognate peptides in *M. murinus* derived samples ($[M+H]^+$ m/z 961 and m/z 1506). The measured mass difference between the proteins from different species (30 Da, Fig. 2) was thus reflected in two tryptic peptides, most likely explained by localisation of amino acid change(s) to one fully digested tryptic peptide and a second peptide resulting from a tryptic miscleavage. A third peptide (m/z 2979 in both species) was accompanied by a +16 Da equivalent (m/z 2995) in both species; this probably contained the putative site of oxidation.

To better characterise the 10 kDa protein, we completed a full protein sequence analysis by mass spectrometry based sequencing *de novo*. The proteins in urine samples containing elevated levels of the 10 kDa from both species were digested with four different proteases – trypsin, endopepeptidase LysC, endopeptidase GluC and endopeptidase AspN to generate peptides that would thus overlap and permit reconstruction of the primary sequence, with the single caveat that the isobaric Leu/Ile amino acid pair could not be discriminated by this method. The protein digests were analysed by LC-MS/MS and the peptide fragmentation spectra were interpreted to yield *de novo* protein sequence data using PEAKS7 software. Initially, high-scoring individual peptide sequences were searched against protein databases using BLAST (http://blast.ncbi.nlm.nih.gov). Several peptide sequences partially matched to a Whey Acidic Protein (WAP) identified in the ring-tailed lemur, *Lemur catta* (Uniprot accession A4K2S4). This protein, full name 'WAP four – disulphide core domain 12 (WFDC12)', was therefore also used to assist the assembly of overlapping peptides (obtained by using endopeptidases of different specificities) for which sequence data was obtained, as a result of which we were able to reconstruct an 87 amino acid sequence (Fig. 4). Sequence analysis of the *M. lehilahytsara* protein also revealed an 87 amino acid protein that was almost identical to the protein from *M. murinus*. However, there was clear evidence for

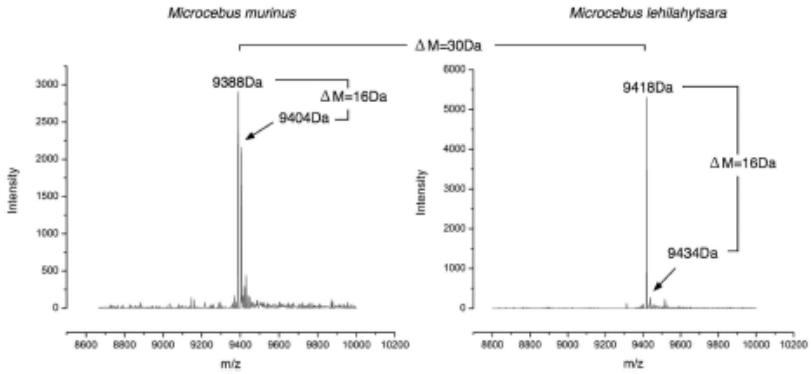


Figure 2. Intact mass analysis of the mouse lemur urinary protein. For urine samples from specific male mouse lemurs within the breeding season, the protein concentration was sufficiently high for analysis of the intact mass of the predominant protein by electrospray ionisation mass spectrometry. Urine samples were analysed for the two mouse lemur species, *M. murinus* and *M. lehilahytsara*, and representative deconvoluted mass spectra, showing the intact mass profiles are presented.

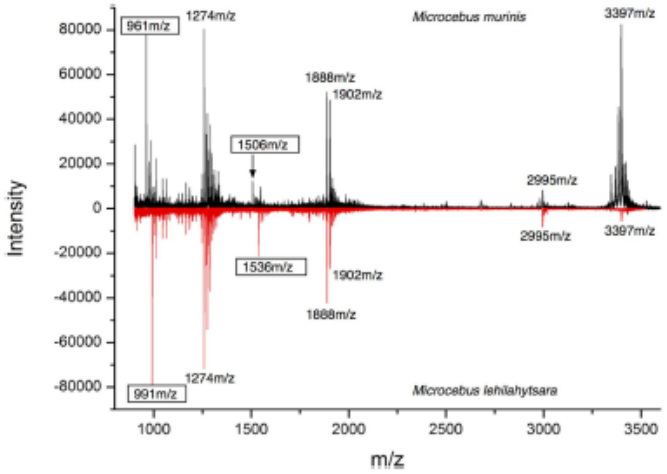


Figure 3. MALDI-TOF mass spectrometry of the mouse lemur urinary protein. For some breeding season males, the predominant 10 kDa band was resolved by SDS-PAGE and the gel fragment was processed by reduction, carbamidomethylation and digestion with trypsin. The resultant peptides were analysed by MALDI-TOF mass spectrometry. Representative spectra are included for *M. murinus* (plotted in a positive direction, black) and *M. lehilahytsara* (plotted in a negative direction for spectral comparison, red). The two boxed peptides exhibited a 30 Da mass difference between the two species.

a single amino substitution in one peptide; corresponding to the singly charged m/z 961 in *M. murinus* and cognate peptide at m/z 991 in *M. lehilahytsara* observed in MALDI-TOF (Fig. 3). The m/z 961 (*M. murinus*) peptide was sequenced as WGNCPAEK, whereas the m/z 991 peptide (*M. lehilahytsara*) was sequenced as WGNCPTEK. This Ala-> Thr substitution accounted for the 30 Da difference in mass. Further, the miscleaved peptide WGNCPAEKGCSIK from *M murinus* would explain the MALDI-TOF peptide at $[M+H]^+$ 1506 m/z that was present as a 30 Da heavier peptide in *M lehilahytsara* (Fig. 3). The protein peak that was evident as 16 Da heavier on the intact mass spectrum in both species was attributable to oxidation of the single methionine residue in the m/z 2979 tryptic peptide (ELGQG**M**APVPQKDTWNVGQVGQEASPQK) leading to a peptide at m/z 2995 (ELGQGM*APVPQKDTWNVGQVGQEASPQK) where * indicates the site of oxidation. This is the only methionyl residue in the protein sequence. Finally, the MALDI-TOF peptide at m/z 1902 is most readily

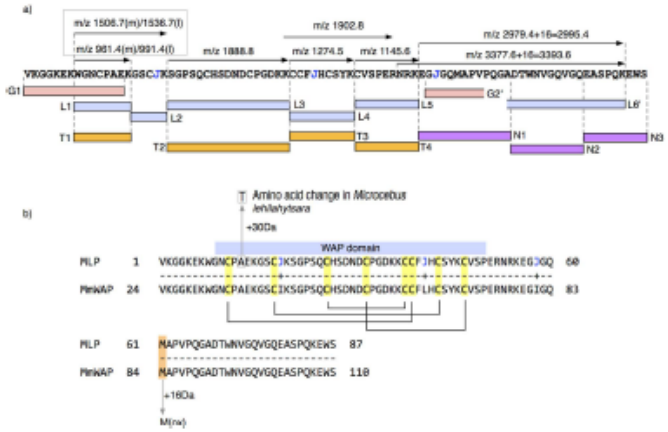


Figure 4. Complete sequencing strategy for the predominant 10 kDa protein. Urine from *M. murinus* and *M. lehilahytsara*, containing high levels of the 10 kDa protein was reduced, carbamidomethylated and digested in solution with four different endoproteolytic enzymes (G: endopeptidase GluC, L: endopeptidase LysC, N:endopeptidase AspN and T: trypsin), the peptides were separated by reversed phase chromatography and sequenced *de novo* by high resolution tandem mass spectrometry. The top panel (**a**) summarises the sequencing strategy (the symbol 'J' is used to define ambiguity between isobaric leucine and isoleucine) The bottom panel (**b**) indicates the homology (MLP: mouse lemur protein sequenced in this study) with the protein inferred from the draft genome sequence for *M. murinus*, (MmWAP, number starting from 24 to accommodate the predicted signal peptide 1–23) assuming the alternate splice variant (see Fig. 5). The putative disulphide bond arrangement observed for other proteins containing the four disulphide core is superimposed on the new protein sequence. Additionally, the position of the single amino change between *M. murinus* and *M. lehilahytsara* and the site of methionine oxidation (for both species) are indicated. The masses of peptides observed by MALDI-TOF mass spectrometry are highlighted at the top of panel a, and the boxed peptide masses confirm the observations from MALDI-TOF mass spectra, where the observed peptides for *M. murinus* (m) or *M. lehilahytsara* (l) are identified.

explained as the sequence KCCFLHCSYKCVSPER (predicted mass 1901.8 Da) that contained four cysteine residues. It is possible that this peptide had evaded reduction and alkylation – the tight disulphide bond 'knot' (see later) could preclude full access to these reagents.

The predicted mass of the newly determined primary protein sequence from *M. murinus* was 9,396 Da, 8 Da higher than the 9,388 Da major peak observed by electrospray ionisation mass spectrometry (Fig. 2). However, a highly conserved feature of WAP domain proteins is the presence of eight cysteinyl residues that form four disulphide bonds[53–55]. The oxidation of eight cysteinyl residues to four disulphide bonds would result in an 8 Da mass decrease, resulting in a predicted mass of 9,388 Da, exactly concordant with the observed mass. Identical arguments for the protein from *M. lehilahytsara* meant that the observed intact mass was also accurately predicted by the protein sequence obtained by sequencing *de novo*, including the single conservative amino acid change (Ala to Thr).

The predicted mass of the protein sequence from *Lemur catta* (Uniprot accession A4K2S4) used to aid assembly of the peptides was 13.2 kDa, significantly larger than the intact mass observed for the mouse lemurs (9.4 kDa). Moreover, there was a region of the *L. catta* protein sequence that was not mapped by any of the peptide *de novo* sequencing data. Failure to acquire evidence for this region of the protein sequence could be for one of two reasons; first, that this region was missing in the mouse lemur sequence or secondly, that this region of the protein has evolved so significantly that homology matching was no longer achievable. The lower mass of the mouse lemur protein favoured the first explanation. When the entire mouse lemur sequence obtained by sequencing *de novo* was used in a BLAST protein search, the strongest match was to a genomic sequence encoding a putative 148 amino acid WAP four-disulfide core domain protein 12 (WFDC12) from *L. catta*. The mouse lemur protein aligned well (75/128 identities, 59%), and all cysteine residues aligned perfectly. However, no sequences aligned to the middle section of the *L. catta* protein (coloured red in Fig. 5a). As the mouse lemur protein was only 9.4 kDa compared to 13.2 kDa for the mature *L. catta* protein, this is corroborative evidence that the corresponding middle section of the sequence of the *M. murinus* protein was absent. Certainly, the overall similarity between the *L. catta* and the *M. murinus* proteins is evident from the dot plot analysis of the two sequences (Fig. 5b). Using the protein sequence described here to search the draft mouse lemur genome (version 1.0) also revealed a match to a putative protein coding gene that also included a nucleotide sequence encoding the 'missing' portion of the grey mouse lemur protein that has been proposed for *L. catta* (Fig. 5a).

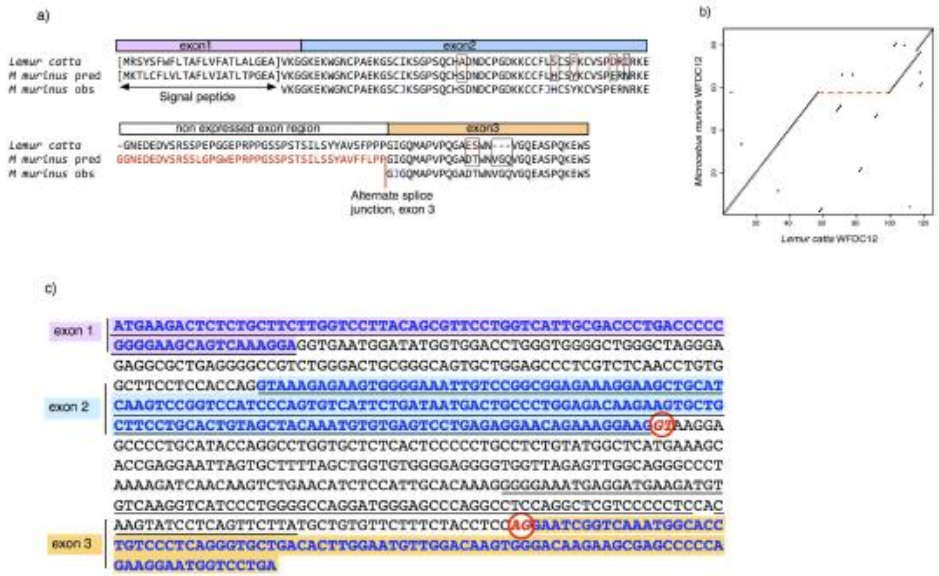


Figure 5. Sequence comparison between mouse lemur and ring tailed lemur. Panel (a) the predicted and observed mouse lemur sequences were aligned with the sequence from the ring tailed lemur (*Lemur catta*, Uniprot accession number A4K2S4). The position of an alternate splice junction in the mouse lemur sequence (http://www.ncbi.nlm.nih.gov/protein/829731644/) is indicated in red. The ambiguity between the isobaric pair leucine and isoleucine is indicated by the letter 'J'. The dot plot (panel b) comparing the mouse lemur and 'long form' predicted ring tailed lemur sequences was generated using the dotPlot routine within the *seqinr* R package (http://seqinr.r-forge.r-project.org/) with a window size of five and scoring matches of three or more amino acids within the window. Panel (c) - Inferred exon/intron structure of *WFDC12* in *Microcebus murinus* genomic sequence of *WFDC12* from Ensembl record ENSMICG00000000986. The four exons from automated prediction are underlined. The three inferred actual exons that encode the WFDC12 protein sequenced in this study are in blue font with coloured background – the translation of these 3 spliced exons is shown as "*M murinus* pred" in panel b. Exons 1 and 2 and intron 1 are the same in both annotations. The predicted 5′ and 3′ splice sites for the newly defined intron 2 are circled and coloured red, with an AG 3′ splice site immediately upstream of exon 3, as expected.

Thus, the predicted sequences from the mouse lemur (and the paralogue from the ring-tailed lemur) genome data were at variance with the observed intact mass analyses from the two species of mouse lemur. Closer examination of the *M. murinus* sequence revealed the potential for an alternative intron/exon structure that would eliminate the coding region for the peptide tract that was not observed (Fig. 5c). The splicing alternatives could thus generate a 'long form' and a 'short form' of the mouse lemur protein. All evidence obtained from urinalysis is consistent with expression and secretion of the short form. Indeed, peptide mapping from multiple individuals to the long form of the protein was compellingly in favour of the short form - in no individual was there evidence for peptides that would be manifest by the long form (Supplementary Fig. S1). Thus, we provide unambiguous evidence for expression of this gene and release of the gene product in male *M. murinus* urine in the breeding season. Evidence of expression of the same 'short form' expression was paralleled in *M. lehilahytsara*. A different prediction algorithm (Gnomon; http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml) did however predict the 'short form'. Despite the high level expression in some breeding season males, the protein sequenced herein matched exactly a predicted protein encoded by a mouse lemur cDNA (accession number XP_012635896; http://www.ncbi.nlm.nih.gov/protein/829731644/). This cDNA was derived from female kidney. Expression in kidney (albeit female) could be consistent with the appearance of this protein in urine. However, it is not possible to infer the site of production of the protein in males specifically within the breeding season.

Confirmatory sequence matching was obtained from in-solution digestion of urinary proteins from multiple individuals. Alignment of all the peptides that were analysed by multiple LC-MS runs provided strong evidence for short form peptides, whereas no peptides corresponding to the additional sequence in the long form were detected (Supplementary Fig. S1). Our data confirm that in urine at least, the 'short form' is expressed, with little evidence for the 'long form'. However, the initial evidence of a 'long form' of this protein derived from the genomic sequence of *L. catta*. To explore this ambiguity further, we were able to acquire urine samples from the communal

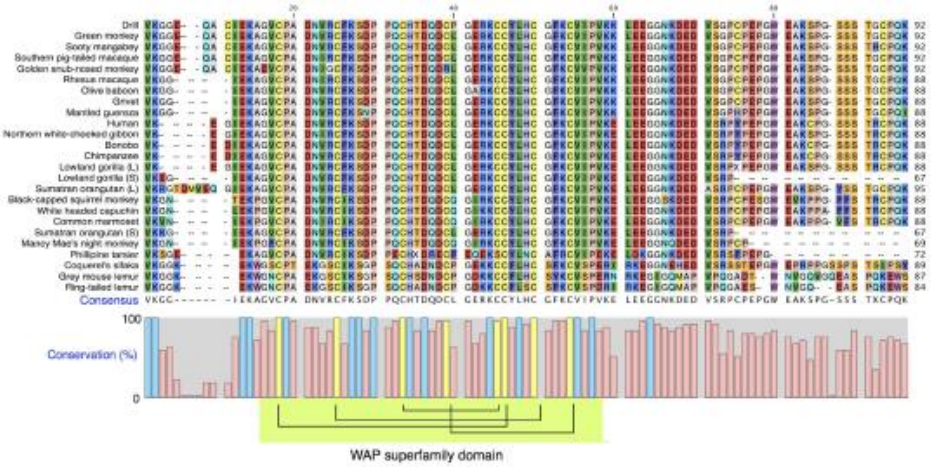


Figure 6. The primate WFDC12 family. All available primate WFDC12 proteins, were aligned using the CLC Sequence Viewer (www.clcbio.com), version 7.6.1. The eight fully conserved cysteine residues are coloured yellow in the conservation graph, in addition to the further 15 residues (blue) within the four disulphide domain that are conserved across all primates shown here.

latrine of 5 male *L. catta* housed at Erlebnis Zoo, Hannover that were collected three times a day, once a week, for three weeks during December and 6 weeks during April-June. We performed creatinine-normalised LC-MS/MS analysis on the tryptic peptides from nine urine samples, matching observed peptides to the predicted long form protein in this species. We additionally performed Glu-C digests on the two samples with the highest abundance of WFDC12, to allow for the large size of a predicted tryptic peptide in the 'long' form, less likely to be seen by tryptic digest. All data was again analysed by PEAKS. As with *M. murinus* and *M. lehilahytsara*, there was no evidence for the long form in this species. We conclude that the short form of the WFDC12 protein is also expressed in *L. catta* urine (Supplementary Fig. S2). As with *M. murinus*, the possibility of an alternate splice site is evident in the *L. catta* sequence (Supplementary Fig. S3).

The mouse lemur urinary protein is a member of the 'whey acidic protein four disulphide core' (WFDC) family, specifically WFDC12. This group of proteins, containing eight characteristically spaced cysteinyl residues that form four disulphide bonds, share limited overall sequence identity, except for the conserved cysteine-rich region and the precise positioning of the cysteine residues[53]. Unlike many other WFDC proteins, WFDC12 is restricted to a short N-terminal segment, the four disulphide core and a C-terminal amino acid section (approx. 40 amino acids) of protein sequence that has no recognised domain structure. BLAST searching of the C-terminal domain *from M. murinus* only returned matches to the *Microcebus murinus* (accession: XP_012635896) and the *L. catta* (accession: A4K2S4) WFDC12 sequences, (the other lemur sequence present in the protein database, from Coquerel's sifaka (accession XP_012513356), did not elicit a match). Indeed, the C-terminal sequence from the sifaka more closely resembled that of other primates. This C-terminal sequence may therefore be specific to the mouse lemurs, a notion that must await further molecular data from other lemur species.

Inferred WFDC12 protein sequences from multiple primates can be readily aligned with the mouse lemur protein (Fig. 6). The position of the eight cysteine residues in the mouse lemur proteins match almost perfectly the arrangement in other primates, confirming the role of the core fold in these proteins (we note that in two sequences, one of the eight cysteine residues has been replaced, but it is not clear if these reflect sequencing uncertainty). In addition, a further 15 residues are fully conserved across all primates from the mature protein N-terminus to 10 residues C-terminal to the last cysteine residue of the four disulphide core. After approximately 10 residues C-terminal to the four disulphide core domain, the sequences diverge very noticeably. For four primates, *Gorilla gorilla* (gorilla; a long and short form are listed in the genomic database), *Pongo abelii* (Sumatran orangutan, also a long and short form), *Aotus nancymaae* (Nancy Ma's night monkey) and *Carlito syrichta* (Philippine tarsier) the protein sequences are predicted to terminate between 10 and 20 amino acid residues from the core. It is possible that these truncated forms reflect uncertainties in early drafts of genome sequences. For the remaining primates a common feature is a C-terminal extension approx. 50 amino acids long.

Within primates, a total of 25 residues at the N-terminal region of the protein, containing the four disulphide core, were completely conserved, suggesting conservation of structure. To gain further insight into the mouse lemur protein, we conducted homology modelling using the Phyre2 package[46]. The disulphide linked core was modelled with confidence (scores > 99), using the known three dimensional structures of other proteins that contain this WFDC domain, including antileukoprotease (2Z7F.PDB) and elafin-like knottins (1FLE.PDB, 2REL.PDB, 1UDK.PDB). The four disulphide core was confidently modelled, and the positions of the eight cysteine

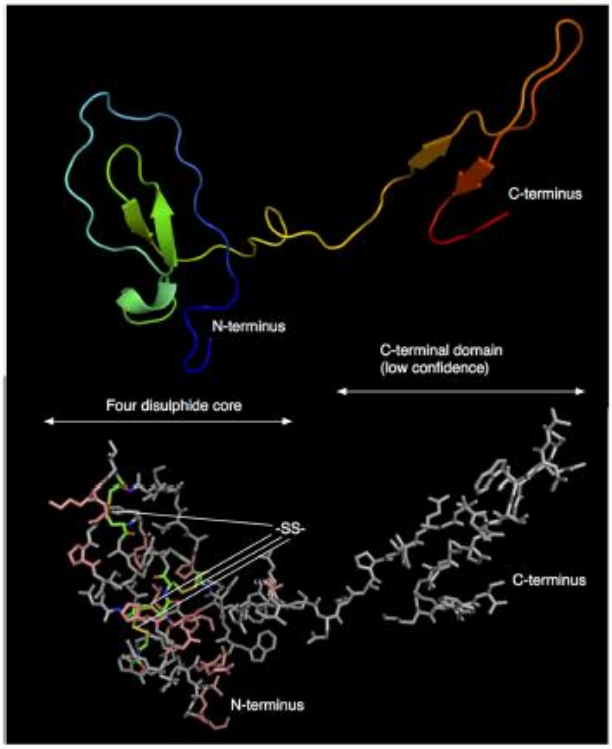


Figure 7. Homology modelling of the mouse lemur urinary protein. The sequence of the mouse lemur protein from *M. murinus* was used to obtain a model structure using the Phyre2 server (http://www.sbg.bio.ic.ac.uk/~phyre2) using the 'intensive' modelling approach. The four disulphide core was modelled to a very level of confidence, but the C-terminal domain cannot be predicted with equivalent confidence.

residues were optimally disposed to generate the four disulphide bonds (Fig. 7). The 15 residues that are conserved in all primate sequence known to date are disposed predominantly towards the exterior of the core fold. The C-terminal domain could not be modelled confidently as there was no specific template against which a structure could be modelled, but it is noteworthy that there is potential for a flexible region between the N- and C-terminal regions.

At present, it is not feasible to make confident assertions about the function of this protein. The high level of expression in males, restricted to the breeding season, is consistent with a role in reproduction or in social interaction during the breeding season. Moreover, not all breeding season males expressed this protein, suggesting that it would be worth exploring a relationship between expression and social hierarchy. Biological functions of some of the WFDC proteins include protease inhibition and antibacterial/antimicrobial activity[56]. The inhibitory loop in WFDC protease inhibitors is not, however, conserved in WFDC12, and preliminary experiments have shown that the mouse lemur WFDC12 is not capable of acting as an inhibitor of pancreatic trypsin or pancreatic elastase (Unsworth, unpublished data). Although the function of this protein in primates is currently unknown, the orthologue of WFDC12 in the mouse (also known as Swam2) has antibacterial activity[56]. Putative antimicrobial activity requires further analysis.

The production of large molecules such as proteins is energetically costly and it can therefore be hypothesized that these molecules are not just "leaking" into the urine after having served physiological functions in the body, but could play a role in chemical signalling in these species. Evidence of positive selection on WFDC12 during primate evolution suggests that this gene may be involved in sexual selection[57]. *WFDC12* transcripts are expressed in the human prostate as well as the skin, lungs and oesophagus[58]. Proteomics expression data (http://pax-db.org/protein/1855755/WFDC12) for human WFDC12 confirms that the protein has been detected in very few tissues (skin, whole organism aggregated data). In skin the protein expression level is at approximately 80 ppm, compared to the urine abundance that exceeds 10^5 ppm. The exceptionally high level of seasonal output of this protein is remarkable. At present, it is not possible to establish the tissue of origin of the mouse lemur protein found in urine, nor is

the tissue distribution of protein expression yet known, although it is tempting to suggest it may be a product of the lemur prostate gland that is activated in the reproductive season. However, the expression of a cDNA encoding sequence in female mouse lemur kidney (see previously) might be suggestive of other tissues of origin. The strong expression in voided urine is consistent with a specialised role rather than being part of the secretion of an active prostate gland; no other prostate-derived proteins are expressed at similar levels in urine (Fig. 1). Mouse lemurs display a behaviour known as 'urine washing', in which voided urine is deliberately deposited on hands, feet and subsequently on the substrate. Moreover, urine contains information about the dominance status of males[40]. It is possible that this protein is part of the scent of a dominant, sexually active male, but such a notion would require further exploration. From the analyses we have conducted thus far, there is no evidence for expression of MUPs in mouse lemur urine. However, it is unlikely that the role of the mouse lemur WFDC12 protein is volatile ligand binding, as it lacks the well-formed, beta barrel enclosed calyx that is a feature of MUP-like proteins. Thus, any role of the urinary WFDC12 protein is unlikely to be mediated by association with low molecular weight ligands.

References

1. Wyatt, T. D. *Pheromones and animal behavior: chemical signals and signatures* (Cambridge University Press, Cambridge, 2013).
2. Liberles, S. D. Mammalian pheromones. *Annu Rev Physiol* **76,** 151–175 (2014).
3. Kaur, A. W. *et al.* Murine pheromone proteins constitute a context-dependent combinatorial code governing multiple social behaviors. *Cell* **157,** 676–688 (2014).
4. Roberts, S. A. *et al.* Darcin: a male pheromone that stimulates female memory and sexual attraction to an individual male's odour. *BMC Biol* **8,** 75 (2010).
5. Roberts, S. A., Davidson, A. J., McLean, L., Beynon, R. J. & Hurst, J. L. Pheromonal induction of spatial learning in mice. *Science* **338,** 1462–1465 (2012).
6. Wilburn, D. B. *et al.* Proteomic and UTR analyses of a rapidly evolving hypervariable family of vertebrate pheromones. *Evolution* **66,** 2227–2239 (2012).
7. Chamero, P. *et al.* Identification of protein pheromones that promote aggressive behaviour. *Nature* **450,** 899–902 (2007).
8. Cheetham, S. A. *et al.* The genetic basis of individual-recognition signals in the mouse. *Curr Biol* **17,** 1771–1777 (2007).
9. Sherborne, A. L. *et al.* The genetic basis of inbreeding avoidance in house mice. *Curr Biol* **17,** 2061–2066 (2007).
10. Hurst, J. L. & Beynon, R. J. Making progress in genetic kin recognition among vertebrates. *J Biol* **9,** 13 (2010).
11. Beynon, R. J. *et al.* The complexity of protein semiochemistry in mammals. *Biochem Soc Trans* **42,** 837–845 (2014).
12. Green, J. P. *et al.* The Genetic Basis of Kin Recognition in a Cooperatively Breeding Mammal. *Curr Biol* **25,** 2631–2641 (2015).
13. Logan, D. W., Marton, T. F. & Stowers, L. Species Specificity in Major Urinary Proteins by Parallel Evolution. *PLOS ONE* **3,** e3280 (2008).
14. Mudge, J. M. *et al.* Dynamic instability of the major urinary protein gene family revealed by genomic and phenotypic comparisons between C57 and 129 strain mice. *Genome Biol* **9,** R91 (2008).
15. Sheehan, M. J. *et al.* Selection on Coding and Regulatory Variation Maintains Individuality in Major Urinary Protein Scent Marks in Wild Mice. *PLoS Genet* **12,** e1005891 (2016).
16. Cheetham, S. A., Smith, A. L., Armstrong, S. D., Beynon, R. J. & Hurst, J. L. Limited variation in the major urinary proteins of laboratory mice. *Physiol Behav* **96,** 253–261 (2009).
17. Gómez-Baena, G., Armstrong, S. D., Phelan, M. M., Hurst, J. L. & Beynon, R. J. The major urinary protein system in the rat. *Biochem Soc Trans* **42,** 886–892 (2014).
18. Phelan, M. M. *et al.* The structure, stability and pheromone binding of the male mouse protein sex pheromone darcin. *PLoS One* **9,** e108415 (2014).
19. Braune, P., Schmidt, S. & Zimmermann, E. Spacing and group coordination in a nocturnal primate, the golden brown mouse lemur (Microcebus ravelobensis): the role of olfactory and acoustic signals. *Behavioral Ecology and Sociobiology* **58,** 587–596 (2005).
20. Buesching, C. D., Heistermann, M., Hodges, J. K. & Zimmermann, E. Multimodal Oestrus Advertisement in a Small Nocturnal Prosimian, Microcebus murinus. *Folia Primatologica* **69,** 295–308 (1998).
21. Glatston, A. R. In *Perspectives in Primate Biology* (ed Seth, P. K.) 63–73 (Today and Tomorrow's Printers and Publishers, New Delhi, 1983).
22. Hohenbrink, P., Dempewolf, S., Zimmermann, E., Mundy, N. I. & Radespiel, U. Functional promiscuity in a mammalian chemosensory system: extensive expression of vomeronasal receptors in the main olfactory epithelium of mouse lemurs. *Frontiers in Neuroanatomy* **8** (2014).
23. Roslinski, D. L., Bhatnagar, K. P., Burrows, A. M. & Smith, T. D. Comparative morphology and histochemistry of glands associated with the vomeronasal organ in humans, mouse lemurs, and voles. *Anatomical Record* **260,** 92–101 (2000).
24. Hohenbrink, P., Radespiel, U. & Mundy, N. I. Pervasive and Ongoing Positive Selection in the Vomeronasal-1 Receptor (V1R) Repertoire of Mouse Lemurs. *Molecular Biology and Evolution* **29,** 3807–3816 (2012).
25. Young, J. M., Massa, H. F., Hsu, L. & Trask, B. J. Extreme variability among mammalian V1R gene families. *Genome Research* **20,** 10–18 (2010).
26. Génin, F. Life in unpredictable environments: First investigation of the natural history of *Microcebus griseorufus*. *International Journal of Primatology* **29,** 303–321 (2008).
27. Hohenbrink, S., Zimmermann, E. & Radespiel, U. Need for Speed: Sexual Maturation Precedes Social Maturation in Gray Mouse Lemurs. *American Journal of Primatology* **77,** 1049–1059 (2015).
28. Perret, M. Environmental and social determinants of sexual function in the male lesser mouse lemur (Microcebus murinus). *Folia Primatologica* **59,** 1–25 (1992).
29. Schilling, A., Perret, M. & Predine, J. Sexual inhibition in a prosimian primate: a pheromone-like effect. *Journal of Endocrinology* **102,** 143–151 (1984).
30. Blanco, M. B. Timely estrus in wild brown mouse lemur females at Ranomafana National Park, southeastern Madagascar. *American Journal of Physical Anthropology* **145,** 311–317 (2011).
31. Eberle, M. & Kappeler, P. M. Selected polyandry: female choice and inter-sexual conflict in a small nocturnal solitary primate (*Microcebus murinus*). *Behavioral Ecology and Sociobiology* **57,** 91–100 (2004).
32. Schmelting, B., Ehresmann, P., Lutermann, H., Randrianambinina, B. & Zimmermann, E. In *Diversité et Endémisme à Madagascar* (eds Lourenço, W. R. & Goodman, S. M.) 165–175 (Société de Biogéographie, Paris, 2000).
33. Eberle, M. & Kappeler, P. M. Sex in the dark: determinants and consequences of mixed male mating tactics in Microcebus murinus, a small solitary nocturnal primate. *Behavioral Ecology and Sociobiology* **57,** 77–90 (2004).
34. Radespiel, U. *et al.* Sexual selection, multiple mating and paternity in grey mouse lemurs, Microcebus murinus. *Animal Behaviour* **63,** 259–268 (2002).
35. Huchard, E., Baniel, A., Schliehe-Diecks, S. & Kappeler, P. M. MHC-disassortative mate choice and inbreeding avoidance in a solitary primate. *Molecular Ecology* **22,** 4071–4086 (2013).
36. Radespiel, U. & Zimmermann, E. The influence of familiarity, age, experience and female mate choice on pregnancies in captive grey mouse lemurs. *Behaviour* **140,** 301–318 (2003).

37. Schwensow, N., Eberle, M. & Sommer, S. Compatibility counts: MHC-associated mate choice in a wild promiscuous primate. *Proceedings of the Royal Society B - Biological Sciences* **275,** 555–564 (2008).
38. Schmid, J. & Kappeler, P. M. Fluctuating sexual dimorphism and differential hibernation by sex in a primate, the gray mouse lemur (Microcebus murinus). *Behavioral Ecology and Sociobiology* **43,** 125–132 (1998).
39. Wrogemann, D., Radespiel, U. & Zimmermann, E. Comparison of reproductive characteristics and changes in body weight between captive populations of rufous and gray mouse lemurs. *International Journal of Primatology* **22,** 91–108 (2001).
40. Perret, M. & Schilling, A. Sexual responses to urinary chemosignals depend on photoperiod in a male primate. *Physiology & Behavoir* **58,** 633–639 (1995).
41. Wrogemann, D. & Zimmermann, E. Aspects of reproduction in the eastern rufous mouse lemur (Microcebus rufus) and their implications for captive management. *Zoo Biology* **20,** 157–167 (2001).
42. Zimmermann, E., Radespiel, U., Mestre-Frances, N. & Verdier, J. M. In *The Dwarf and\ Mouse Lemurs of Madagascar* (eds Lehman, S. M., Radespiel, U., Zimmermann, E., Mestre-Frances, N. & Verdier, J. M.) 174–194 (Cambridge University Press, Ca,bridge, 2016).
43. Radespiel, U., Ehresmann, P. & Zimmermann, E. Species-specific usage of sleeping sites in two sympatric mouse lemur species (Microcebus murinus and M. ravelobensis) in northwestern Madagascar. *Am J Primatol* **59,** 139–151 (2003).
44. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Sci* **67,** 1–48 (2015).
45. Zhang, J. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* **11,** M111.010587 (2012).
46. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10,** 845–858 (2015).
47. Han, X., He, L., Xin, L., Shan, B. & Ma, B. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *J Proteome Res* **10,** 2930–2936 (2011).
48. Han, Y., Ma, B. & Zhang, K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J Bioinform Comput Biol* **3,** 697–716 (2005).
49. Vizcaíno, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R. & Hermjakob, H. Update of the PRIDE database and related tools. *Nucleic Acids Res* **44**(D1), D447–D456 (2016).
50. Beynon, R. J. *et al.* Polymorphism in major urinary proteins: molecular heterogeneity in a wild mouse population. *J Chem Ecol* **28,** 1429–1446 (2002).
51. Robertson, D. H., Hurst, J. L., Bolgar, M. S., Gaskell, S. J. & Beynon, R. J. Molecular heterogeneity of urinary proteins in wild house mouse populations. *Rapid Commun Mass Spectrom* **11,** 786–790 (1997).
52. Robertson, D. H., Hurst, J. L., Searle, J. B., Gündüz, I. & Beynon, R. J. Characterization and comparison of major urinary proteins from the house mouse, Mus musculus domesticus, and the aboriginal mouse, Mus macedonicus. *J Chem Ecol* **33,** 613–630 (2007).
53. Bingle, C. D. Towards defining the complement of mammalian WFDC-domain-containing proteins. *Biochem Soc Trans* **39,** 1393–1397 (2011).
54. Ranganathan, S., Simpson, K. J., Shaw, D. C. & Nicholas, K. R. The whey acidic protein family: a new signature motif and three-dimensional structure by comparative modeling. *J Mol Graph Model* **17,** 106–13, 134 (1999).
55. Simpson, K. J. *et al.* The gene for a novel member of the whey acidic protein family encodes three four-disulfide core domains and is asynchronously expressed during lactation. *J Biol Chem* **275,** 23074–23081 (2000).
56. Hagiwara, K. *et al.* Mouse SWAM1 and SWAM2 are antibacterial proteins composed of a single whey acidic protein motif. *J Immunol* **170,** 1973–1979 (2003).
57. Hurle, B., Swanson, W., Nisc, C. S. P. & Green, E. D. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. *Genome Res* **17,** 276–286 (2007).
58. Lundwall, A. & Clauss, A. Identification of a novel protease inhibitor gene that is highly expressed in the prostate. *Biochem Biophys Res Commun* **290,** 452–456 (2002).

Acknowledgements

J.U. is grateful to the Biological Sciences and Biotechnology Research Council for a PhD studentship (BB/F017502/1). We are grateful to Dr Philip Brownridge for excellent instrumentation support in the Centre for Proteome Research. We thank Sarah Hohenbrink for helping with the urine sampling of the mouse lemurs at the Institute of Zoology, University of Veterinary Medicine Hannover, and the Hannover Zoo for providing the urine samples of *L. catta*. We furthermore thank Achim Sauer, Wolfgang Mehl, Lisabelle Früh, Elisabeth Engelke and Birgit Haßfurther for efficient caretaking and handling routines at the breeding facility of the Institute of Zoology, TiHo Hannover.

Author Contributions

U.R., R.J.B., E.Z. and J.L.H. conceived the idea. U.R. and E.Z. designed a protocol to collect the lemur urine samples. J.U., G.L. and A.J.D. conducted the analyses on urinary proteins. R.J.B., J.L.H., J.U., G.L. and G.G.B. were responsible for analysis of mass spectrometric and urinary protein data. All authors contributed to writing of the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Unsworth, J. *et al.* Characterisation of urinary WFDC12 in small nocturnal basal primates, mouse lemurs (*Microcebus spp.*). *Sci. Rep.* **7,** 42940; doi: 10.1038/srep42940 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.